EVOLUTION OF THE RIBOSOMAL COMMON CORE

A Dissertation
Presented to
The Academic Faculty

By

Chad R. Bernier

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Chemistry and Biochemistry

Georgia Institute of Technology
December 2014

EVOLUTION OF THE RIBOSOMAL COMMON CORE

Approved by:

Dr. Loren Williams, Advisor
School of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Roger Wartell
School of Biology
Georgia Institute of Technology

Dr. Steve Harvey
School of Biology
Georgia Institute of Technology

Dr. Adegboyega Oyelere
Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Nicholas Hud
School of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Eric Gaucher
School of Biology
Georgia Institute of Technology

Date Approved: November 14, 2014

ACKNOWLEDGEMENTS

First, my deepest gratitude is to my advisor, Dr. Loren Williams, for his long-term support, patience, professional guidance, and encouragement. My thanks also go to Dr. Nicholas Hud, Dr. Steve Harvey, Dr. Roger Wartell, Dr. Adegboyega Oyelere, and Dr. Eric Gaucher for serving on my thesis committee despite their extremely busy schedules. Their valuable insight was always appreciated.

Special thanks go to Dr. Chiaolong Hsiao, Dr. Derrick Watkins, and Dr. Anton Petrov. They have been exceptional mentors and friends, always there for me, for both professional and personal concerns. In addition to mentorship and friendship, Anton made this whole project possible. This project is more than any one person could do.

I acknowledge all other members and alumni of Dr. William's group, especially Caitlin Prickett, Jessica Bowman, Eric O'Neill, Kathryn Lanier, and Nicholas Kovacs. Jessica kept everything running and was always able to provide special assistance when requested. Eric was especially helpful late at night when everyone else was gone and was an excellent coworker to discuss anything with. Kathryn has been especially helpful at the end, ensuring that my defense went over smoothly. She is also working hard at making Anton and I's model a physical reality. Nicholas is a wonderful friend, roommate, and mentee. Nicholas is working hard expanding the model and our database to include ribosomal proteins.

Finally, I'd like to thank all my family and friends who have been very supportive. I especially thank Guillermo Alas and my parents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, LIST OF ABBREVIATIONS

1D	one dimensional
2D	two dimensional
3D	three dimensional
Å	Angstrom
AES	Ancestral Expansion Segment
API	Application Programming Interface
CADS	Conservational Analysis DataSet
CRW	Compartitive RNA Website
cryo-	
EM	electron cryomicroscopy
CSV	comma separate values
CT	Connectivity Table
cWW	cis Watson / Watson
DARC	Database of Aligned Ribosomal Complexes
DCC	decoding center
DNA	deoxyribose nucleic acid
EPS	encapsulated PostScript
FR3D	Find RNA 3D
GTP	Guanosine-5'-triphosphate
GUI	Graphical User Interface
ITS2	Internally Transcribed Spacer 2
LSU	Large SubUnit

LUCA	Last Universal Common Ancestor
Mg	Magnesium
mRNA	messenger RNA
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information
NDB	Nucleic acid DataBase
NMR	Nuclear Magnetic Resonance
PDB	Protein DataBank
PDF	Portable Document Format
PNG	Portable Network Graphics
PTC	Peptidyl Transferase Center
RMSD	Root Mean Squared Deviation
RNA	Ribonucleic acid
rProtein s	Ribosomal Proteins
rRNA	ribosomal RNA
SSU	Small SubUnit
SVG	Scalable Vector Graphics
tRNA	transfer RNA
VARNA	Visualization Applet for RNA

SUMMARY

In 1967, Carl Woese chose the translation system as his platform for addressing some of the deepest questions in biology. Woese succeeded because translation is indeed a window to primeval events, molecular structures, and chemical processes that directed the broad course of life on Earth. In 2014, we know that translation is a unique province in the biological world. The translational system transforms information from polynucleotide to polypeptide in a spectacular molecular choreography revealed by high-resolution ribosome structures from all three domains of life and by a massive and ever-expanding sequence database. The translation system retains an interpretable molecular record of biology from before the last universal common ancestor (LUCA) and is our guide to the world of primordial molecules.

Understanding the origin of life requires understanding the origin of translation, which in turn, requires understanding the origin of the ribosome. Ribosomes are complex structures consisting of hundreds of thousands of atoms. Here, we describe how we organized ribosomal information into a high-quality database. We also describe a new visualization webapp, RiboVision.

RiboZones and RiboVision are productivity tools that lower the learning curve for ribosomal research. RiboZones makes the ribosome more accessible. RiboVision especially helps create beautiful publication ready figures in a fraction of the labor and time previously required. It is only through the creation of RiboZones and RiboVision through which the rest of this dissertation became feasible.

We constructed a high-quality sequence alignment of ribosomal sequences for both the LSU and the SSU rRNA. Each ribosomal sequence is complete, allowing detailed, low background statistics to be computed. The sequence alignment broadly samples the tree of life according to available data. The alignment was adjusted for maximum agreement with 3D superimpositions of multiple ribosomal structures.

We defined a nucleotide-level definition of the common core of the ribosome, as the RNA that is present in 95% of the sequences in our alignment. Multiple versions of the common core were created, including the universal common core, the prokaryotic common core, and domain specific common cores. The definition allows statistics to be computed for various use-cases. For example, with RiboVision visualization technology, it is possible to see which helices are optional, in which of the three domains of life, and what the minimum helical length is for each helix. The common core represents universal ribosomal function for all extant life and can be exploited to learn about ribosomal function and structure.

We discovered that ribosomal RNA grows mostly be helix extension and helix insertion. When a helix is inserted, it minimally perturbs the underlying helix. We call this pattern *'insertion fingerprints'*. Insertion fingerprints are found throughout the common core and the eukaryotic expansion segments.

Insertion fingerprints were used to divide the ribosomal RNA into units called ancestral expansion segments (AES's). AES's make ideal structural, functional, and evolutionary units. The AES's are arranged into the first complete experimentally testable model of ribosomal evolution. The model can be refined over time as new information is discovered.

CHAPTER 1

INTRODUCTION

The ribosome is at the center of the translation system, a key component to life. The genetic code, which is nearly universally conserved among all terrestrial life with a few specific exceptions, is embodied in modern tRNA and aminoacyl-tRNA synthetases. However, it is the ribosome which ensures the correct translation between the mRNA genetic message and the finished protein product. Since the genetic code is nearly universal and the ribosome must predate the genetic code, the origin of the ribosome is rooted before the Last Universal Common Ancestor (LUCA).[1] This makes every species alive a molecular fossil for determining the origin of the ribosome.

Atomic-resolution three dimensional structure models, such as those resolved by NMR and X-ray crystallography, contain a vast amount of information within themselves. It is a challenge to fully utilize all this information, especially in large structures, such as the ribosome. The bacterial ribosome is composed of two macromolecular units, the small (30S) and large (50S) subunits. Each subunit is composed primarily of ribosomal RNA (16S and 23S in the small and large subunits respectively) and accessory proteins. Atomic-resolution x-ray structures (<3.0Å) of ribosomes are available for several species, including, *Escherichia coli* (bacteria), *Haloarcula marismortui* (archaea, SSU only), and *Saccharomyces cerevisiae* (eukaryotes). All atom 3D structure models, based on high resolution cryo-EM data are available for *Drosophila melanogaster* and *Homo sapiens*. The 3D structures confirm that all ribosomes share a structural common core, which composes over 90% of prokaryotic ribosomes. Therefore, LUCA already had a fully functional ribosome.

1.1. The Origin of the Ribosome

The origin of life is the biggest unanswered question in biology. The origin of life and the origin of the ribosome are intricately linked. Ribosomes are a defining feature of life and are precisely at the interface between chemistry and biology. The ribosome coevolved with the fundamental rules of biology.

The modern ribosome is an extremely complex machine composed of highly specific RNA (rRNA) and highly specific proteins (rProteins). In order to function, other highly specific components made out of different RNA and different proteins need to be functional. Many antibiotics and poisons work by interfering with ribosomal function. Such a system could not evolve by pure chance. It is a classic chicken and egg paradox.

Understanding the origin of the ribosome can be achieved by understanding its structure and function. Early forms of life were not as dependent on ribosomal regulation, ribosomal speed, or ribosomal accuracy and would have used simpler ribosomes. Since there are no partially developed ribosomes still existing, they must be modeled with computers and tested in the laboratory. Understanding the ribosome's structure and function is crucial to building such evolutionary models.

The ribosome presents extraordinary challenges (and opportunities) for data interpretation, visualization, analysis, and management. The ribosome is unique in its combination of importance, size, structural/functional complexity, and enormity in available information. Information available on ribosomes includes: a) atomic positions of rRNA, tRNA, mRNA, rProteins, translation factors, associated ions, water molecules and antibiotics, obtained from x-ray, cryo-EM and NMR structures, b) very large and rapidly growing databases of rRNA and rProtein sequences, c) molecular interactions

among ribosomal components, inferred from 3D structures, phylogeny, or activity assays, d) phylogenetic relationships and mutational patterns of various ribosomal components, e) sites and types of modification of rRNAs and rProteins, f) chemical mapping and reactivity data, and g) functional and dynamical information.

1.2 Data Visualization

Organizing and visualizing the amount of data available and required is challenging, taking up a lot of time and labor. Visualizing the ribosome requires use of both 2D and 3D structure representations. Manually mapping information onto both 2D and 3D structures was severely limiting the types of analysis we could contemplate. We required integration of structure, function, and phylogeny represented in 1D, 2D and 3D, incorporating a broad variety of data types. No databases or software contained a complete set of data, a complete analysis suite, or a complete visualization suite.

Studying the evolution of the ribosome requires studying multiple species at the same time. More accurately, it requires studying multiple versions of purposely-incomplete ribosomal structures from multiple species. An "experiment" consists of generating multiple versions of figures and comparing them. Figure generation could not be a bottleneck, many processes needed to be automated.

1.3 Using RiboZones to build an evolutionary model

As we collected, organized, and fixed problems with data, we realized that our dataset would be helpful to a larger scientific community. The same is true about our software tools that we developed. We decided to be open with all our data and our software. We designed our software with maximum interoperability, flexibility, and

expandability in mind. The software makes use of abstraction layers and open standard file formats. We call our collection of data and software RiboZones.

This dissertation uses RiboZones to build a model of ribosomal evolution. Chapter 2 is background information. In Chapter 3, the major components of RiboZones are described. Chapter 4 describes how RiboZones was used to make an accurate structure-based multiple sequence alignment. Chapter 5 uses the alignment to make the first rigid per-nucleotide definition of the common core. Chapter 6 is where everything comes together. We observe how the ribosome has grown from the common core, leading to a new way of thinking about the structure of the ribosome. The ribosome can be broken into structural pieces, called ancestral expansion segments (AESs') which have functional and evolutionary significance. The AES's may someday inspire a reorganization of the way secondary structures are drawn. AES's can be arranged into a whole family of evolutionary models. One version of these models is presented. Using RiboZones, researchers can design specific experiments to test either the standard RiboZones model, an alternative model, or any tangential hypothesis about the ribosome. RiboZones is designed for both computational /theoretical scientists and experimental scientists, as these users tend to have different needs and different skill sets.

Finally, Chapter 7 is a general discussion of the main components of RiboZones. First, the usefulness of future of RiboZones, particularly, the released and published component, RiboVision is discussed. Second, our sequence alignment is discussed, and how it stands out from other alignments for future work. Third, the potential of the common core is discussed. Finally, the usefulness and potential of our model is discussed.

CHAPTER 2

LITERATURE REVIEW

2.1 Background

The ribosome performs all coded synthesis of proteins in every cell. The ribosome is composed of two macromolecular assemblies, called the large subunit (LSU) and the small subunit (SSU). Each subunit is composed of a number of ribosomal RNA molecules (rRNA) and ribosomal proteins (rProteins). The primary function of the LSU is to catalyze formation of a new peptide bond between an amino acid, and a growing protein chain. The LSU does this through careful positioning of two charged transfer RNA molecules (tRNA).[2,3] The primary function of the SSU is to read the messenger RNA (mRNA), and only allow the correct tRNA into the LSU. Once the new peptide bond is formed, the ribosome moves down the mRNA, through a process called translocation, and repeats the process for the next amino acid, until the protein is complete.[4,5]

All ribosomes share a similar structure,[6] which is illustrated using *Escherichia coli* as an example. The LSU secondary structure is shown in **Figure 2.1A.** The LSU tertiary structure is shown in **Figure 2.1B.** The SSU secondary structure is shown in **Figure 2.1C.** The SSU tertiary structure is shown in **Figure 2.1D.** A fully assembled ribosome (**Figure 2.1E)** contains an mRNA molecule, and three tRNA molecules. The mRNA molecule in this structure is buried inside and not visible from the chosen angle. A cartoon of an mRNA is drawn for illustration of approximate placement. There is no growing peptide chain in this structure, so a cartoon protein was drawn. The new protein

chain would come out the exit tunnel of the LSU, which would be on the top of the assembled ribosome from this view.

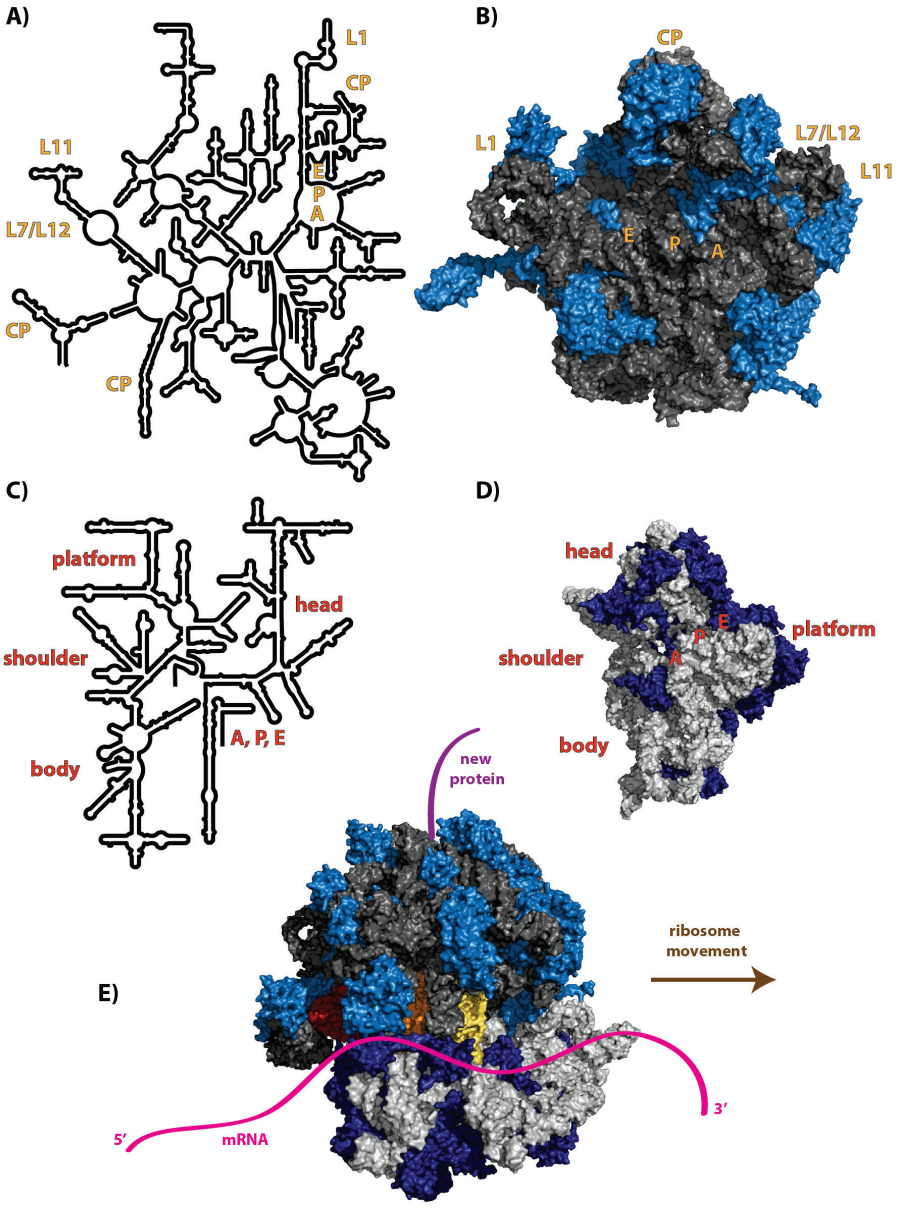

Figure 2.1. Overview of ribosome structure, using *E. coli*. A). Secondary structure of the LSU. B). 3D structure of the LSU. RNA is dark gray, rProteins are light blue. C). Secondary structure of the SSU. D). 3D structure of the SSU. RNA is light gray, rProteins are dark blue. E) Assembled ribosome. A-site tRNA (yellow), P-site tRNA (orange), and E-site tRNA (red). A cartoon mRNA (pink) and cartoon new protein (purple) are drawn in their approximate positions.

Three tRNA molecules are in the three "sites" of the ribosome. The A-site holds the incoming aminoacyl-tRNA (aa-tRNA). The SSU only allows the correct aa-tRNA into the A-site when functioning properly. Once inside, the tip of the A-site tRNA is placed very closely to the tip of the P-site peptidyl-tRNA. The P-site tRNA is attached to the growing protein chain. The ribosome transfers the protein chain from the P-site tRNA to the A-site tRNA. The now deacylated P-site tRNA moves into the E-site, displacing the previously deacylated tRNA in the E-site. The previously A-site tRNA, which now has the protein chain attached to it, becomes the new P-site tRNA. The elongation cycle repeats when another correct aa-tRNA is allowed into the LSU. The processes for protein initiation and termination are also well studied.[7-13]

For the LSU, the A, P, and E sites, are composed of rRNA in Domain V, and collectively called the peptidyl transferase center (PTC). In the classic orientation, called crown view (**Figure 2.1B**), the PTC is approximately in the center. There are three projections at the top. The left projection is the L1 protuberance, which binds to rProtein L1. L1 is involved in translocation, helping to move the deacylated P-site tRNA into the E-site. The middle projection is called the central protuberance (CP). It is composed of rRNA from several domains, including part of the 5S rRNA, plus several rProteins. The right projection is called the L7/L12 stalk, because it binds to L7 and L12. Together with other rProteins, the L7/L12 stalk forms an arm that helps position tRNA and elongation factors. It also forms the critically important GTPase center, for the GPTase domains of the various elongation factors to bind to, allowing coupling of ribosomal movements to the energy provided from hydrolyzing GTP. This energy coupling allows the ribosome to reach catalytic rates appropriate for supporting modern life.[14]

For the SSU, the A, P, and E sites, are composed of rRNA in the 3'minor domain and collectively called the decoding center (DCC). The structure of the SSU (**Figure 2.1C)** is organized into a head, body, shoulder, and platform. The mRNA enters through the shoulder. The tRNAs bind and move across the platform. The head can rotate slightly, participating in translocation. The body provides structure and secures the SSU to the LSU.

The role of ribosomal proteins are numerous and not completely known. They are heavily involved in proper biological assembly.[15] Most rProteins are not essential for peptidyl transferase activity.[16,17] L2 is particularly important.[18-23] The primary role of rProteins is to stabilize rRNA and support ribosomal functions. However, rProteins have gained additional functions.[24] They are even involved in regulation, including of themselves.[25] This dissertation focuses on the rRNA, but RiboZones in general supports rProteins.

2.2 Ribosomal Sequences and Structures

Ribosomal RNA sequences are obtained through sequencing of the genes (DNA) used to transcribe the rRNA. It is rare for RNA to be directly sequenced. For most species, the rRNA gene is sequenced specifically.[26,27] Sometimes, sequences are obtained through whole genome sequencing. [28,29] The primary databases for storing DNA sequences are GenBank[30] and The European Nucleotide Archive.[31] There are also more specialized databases, such as GeneDB,[32] which cover only a subset of species. Sequences of interest can be obtained through either searching for gene annotations, or with tools such as BLAST[33] and BLAT.[34]

Databases specific to ribosomal RNA are also available. The major databases are RDP-II[35], SILVA[36], and GreenGenes.[37] SILVA is the only database available which covers, the three domains of life, bacteria, archaea, and eukarya. SILVA is also the only major database which covers both the LSU and SSU.

The CRW site is the only major database of secondary structures.[38] Otherwise, secondary structures are determined or predicted independently.[39-42] Williams et al made all the secondary structures in this dissertation based on our published methodology.[43,44]

Tertiary structures are primarily obtained from the Protein Data Bank (PDB). [45] More specialized databases get the initial structures from the PDB. These include the Nucleic Acid DataBase[46] (NDB), the DARC site: a database of aligned ribosomal complexes,[47] and RNA 3D Hub.[48] The specific structures used are described in **Chapter 3.2.1**

2.3 Structure and Sequence Alignments

The DARC site: a database of aligned ribosomal complexes,[47] was the only database which superimposes ribosomal structures onto a common reference frame. Since ribosomes are complex, and the PDB format is old, alignment of whole ribosomes is non trivial. Ribosomes come in at least 2, possibly 3, or 4 PDB files each. Aligning the molecules in all 4 PDB files at once is challenging for non-experts. Unfortunately, DARC hasn't been updated since 2012.

In practice, there are two de facto standard reference rRNA alignments: Gutell's CRW alignment,[38] and SILVA's seed alignment.[36] These are described in **Chapter 4.1**. Alignments for the SSU only are available from RDP[35] and GreenGenes.[37] Otherwise, researchers generally make their own alignments through a variety of methods. [42,49-51]

CHAPTER 3

CONSTRUCTION OF RIBOZONES

3.1 Motivation for RiboZones

The ribosome contains a large amount of information packed into a large molecule. Studying a molecule with hundreds of thousands of atoms presents technical, scientific, and visual challenges. The ribosome with its large size, unusual folding patterns, unusual base pairing patterns, and interactions with ions and rProteins, presents unique challenges. Most software released to study RNA is ill suited for studying the ribosome. The ribosome needs specialized software.

The ribosome presents extraordinary challenges (and opportunities) for data interpretation, visualization, analysis, and management. To understand the role of the ribosome in the origin of life, or in any other context, we are developing a web portal (RiboVision[52]) for ribosomal data management and representation. The ribosome is unique in its combination of importance, size, structural/functional complexity, and enormity in available information. We consider the ribosome to be the most information-dense space in the known universe.

3.1.1 Early RiboZones history

Traditionally, when a researcher wants to integrate data from multiple sources in a program and analyze it in a new method, they need to write their own software to do so. That is also how RiboZones began. At first, before it was named, RiboZones was a 3D structure analysis program written in MATLAB. Later, a sequence analysis program was developed. The programs took advantage of object-oriented techniques and shared many

of the same classes. Eventually, they gained support for writing output files in PyMOL script format, so results could be directly visualized in PyMOL.

The PyMOL script support inspired creation of the RiboZones PyMOL script collection. Instead of an unorganized collection of standalone PyMOL scripts, why not a hierarchal collection of module PyMOL scripts? Just like with computer code, this increases maintainability, compatibility, and usability, and reduces development time. Combined with pre-superimposed structures, these features really help comparisons between species.

3.1.2 Secondary structure bottleneck

Visualizing results on secondary structures quickly became the bottleneck. At that point, we visualized results by placing and coloring circles, by hand, one by one, onto the secondary structure in Adobe Illustrator. This is a very time-consuming labor intense process, and extremely limited the analysis we could perform. Sometimes results were drawn by hand with colored markers to see if the results were significant enough to draw on the computer. Some research groups have claimed to use something like a python or perl script to draw secondary structures, but they never released their scripts.

We found a way to process the secondary structure files into MATLAB, through SVG files. From there, we developed code to have MATLAB draw the circles and lines, which would perfectly superimpose onto the secondary structure. Scripts were written for Adobe Illustrator using their JavaScript API to process the files as drawn by MATLAB.

Removing the secondary structure bottleneck enabled research that is more diverse and vastly increased productivity. We realized how useful this could be to a wider scientific community and wanted everyone to use it. We decided that making a webapp

version of our visualization software would be the best way to distribute it and have the widest appeal. RiboVision was born, and the rest of our data and software tools were integrated together around RiboVision.

3.1.3 RiboZones Philosophy

RiboZones is intended to be a general purpose set of data and tools for a wide variety of scientists and students. The design goals include:

- Integration of multiple data sets from a wide variety of sources
- Simultaneous viewing of data in 1D, 2D, and 3D representations
- Interactive data and structure viewing
- Standard input and output formats
- Ease of use
- Good documentation
- Community support
- Open source and open data

RiboZones is already useful, but has much greater potential, if the community accepts it. RiboZones cannot replace specialized databases, serious scientific software, traditional publishing, etc. RiboZones supplements these things, acting as a glue which integrates them together. If everyone made their data and software compatible with RiboZones, it would increase the productivity of everyone. It would also increase the likelihood of their data or software being used and cited. There is currently far more data available than the world's processing capability. The more accessible data will be processed first.

3.2 Data Collection and Organization

3.2.1 3D Structures

Three-dimensional structures of ribosomal particles were obtained from the PDB database. The x-ray structure of Steitz [53] was used for *H. marismortui* (PDB entry 1JJ2, resolution 2.4 Å). The x-ray structure of Ramakrishnan [53,54] was used for *T. thermophilus* (PDB entries 2J01, 2J00, resolution 2.8 Å).The x-ray structure of Cate [55] was used for *E. coli* (PDB entries 3R8S, 4GD1, resolution 3.0 Å), and the x-ray structure of Yusupov [56] was used for *S. cerevisiae* (PDB entries 3U5B, 3U5C, 3U5D, 3U5E, resolution 3 Å). Ribosomes of *D. melanogaster* and *H. sapiens* are the cryo-EM structures of Beckmann [57] (PDB entries (3J38, 3J3C, 3J39, 3J3E for *D. melanogaster*, resolution 6 Å; PDB entries 3J3A, 3J3B, 3J3D, 3J3F, resolution 5 Å for *H. sapiens*).

Initially, all ribosomes were structurally aligned into the coordinate space of the processed structures from Williams.[58] PyMOL's cealign feature was used to either align by L2 only, or align by the whole LSU rRNA. In Chapter 4, DARC aligned structures are used.[47] Having all structures structurally aligned has been helpful.

3.2.2 2D Structures

Secondary structures are instrumental in RiboZones. Secondary structures have been constructed to match each of the 3D structures above. Secondary structures are initially constructed in XRNA.[59] New secondary structures are built by using a previous secondary structure and the sequence alignment to do templating. VARNA might be used if it supported templating. [60] Templating allows the RiboZones secondary structures to be as homologous and superimposable as possible, facilitating both comparative analysis and figure generation.

Secondary structures are stored in several formats. Structures are initially processed from native XRNA files when available. RiboLab software (MATLAB) has an internal storage format for interoperation of different analysis programs. There are utilities to convert to CT files as requested. RiboVision has an external format (CSV based) designed to be used for preparing new datasets for RiboVision. For visualizing structures, SVG files are generated. SVG is easily convertible into other vector or bitmapped image formats. Software is available to help interconvert all formats. The release of RiboZones will include all structures available in all formats. Everything is designed for open and convenient data access.

3.2.3 Primary Structures

There are thousands of primary sequences available across most of the tree of life. A smaller subset was desired to make high quality data collection more manageable. Currently, the species set contains 133 organisms, 67 bacteria, 36 archaea, and 30 eukaryotes. Taxonomic ID's have been collected to facilitate partially automated sequence searching.

For each organism, LSU and the SSU sequences have been compiled, along with sequences for several ribosomal proteins. Sequences were primarily taken from SILVA and NCBI databases. These sequences were supplemented from other sources. Some sequences are combinations of partial sequences from multiple sources. As gene annotations were often not sufficient, contigs and chromosomal assemblies of DNA were examined to find missing pieces of rRNA. The exact 3' and 5' ends of rRNA had to be approximated based on observed patterns. For species with fragmented rRNA, all fragments were combined into the correct order.

3.2.4 Base Pair Interactions

Nucleotide-nucleotide interactions are complex and numerous. RiboZones is designed for multipurpose use. It is not known what kinds of interactions users will want to study, so a wide variety of interactions is required. Rather than implement software to do such calculations, it is outsourced. RNA-RNA interactions were collected from RNA 3D Hub.[48] The "FR3D all pairwise interaction annotations" were copied from the website for each structure available in RiboZones. RiboLab can parse these files and convert them to its internal base-pairing format. RiboLab collects other types of interactions not supported by RNA 3D Hub and stores them in the same base-pairing format. Base pairs can be filtered by a wide variety of criteria and visualized in direct EPS files, or converted into RiboVision format.

3.2.5 Other RNA Interactions

RiboLab has tools to calculate a variety of molecular interactions. In addition to supporting custom calculated interactions, a few standard ones are precalculated. Most notably, protein and magnesium contacts and interactions are calculated. Protein contacts are generally calculated as interatomic distances between amino acid and nucleic acid atoms of 3.4Å or less. Protein interactions are defined as a single amino acid contacting more than one nucleotide. Magnesium contacts are generally calculated as interatomic distances between magnesium ions and nucleic acid atoms of 2.4Å or less. Magnesium interactions are defined as a single magnesium ion contacting more than one nucleotide.

3.2.6 Organization

There is a huge amount of information known about the ribosome in the literature, with more being discovered every year. Organizing all this information is an ambitious

goal that would require the cooperation and participation of the entire ribosomal community. RiboZones is providing the infrastructure to organize data from a wide variety of sources. An overall goal is to be able to easily interface with even more specialized databases already existing and allow easy combination of multiple types of data. RiboZones is a unified analysis suite that nicely supplements tools that are more specialized. RiboZones provides a starter set of data for students and researchers new to the field. A common theme is rapid hypothesis testing. Preliminary analysis can be performed quickly to determine if a more detailed analysis is justified. The gaps between structural biologists, experimental biologists, and bioinformaticians are bridged.

3.2.7 Quality

Data is checked for consistency. When integrating data from several sources there are often slight errors, or at least inconsistencies. These present problems to automated data analysis algorithms and researchers' understanding. The perfect example is *Thermus thermophilus* LSU rRNA sequence. The X-ray crystal structure of *T. thermophilus* HB8, published in 2006, [54] was the first whole 70S ribosome with a resolution under 3Å. The researchers chose to use an *E. coli* numbering scheme in the crystal structure. While this sounds like a good idea from the perspective of a structural biologist analyzing the structure, it doesn't work well for everyone else.

First, there are no tools designed to deal with an alternative numbering scheme. Sequence level data will use the natural numbering system. Researchers are forced to individually produce a mapping file between the natural numbering system and the *E. coli* numbering system. When they are looking at the sequence in a computer program, it will use the natural numbering system. Similarly, when they are measuring nucleotide

level data in the lab, it will use the natural numbering system. This complicates analysis greatly and causes confusion at publication time.

The desire to use secondary structures makes the problem worse. Most people would obtain the secondary structure from Noller's Center for Molecular Biology of RNA.[61] There, a secondary structure of *T. thermophilus* rRNA is provided. Thankfully, it uses *E. coli* numbering too. However, there are several issues. The actual RNA sequence used is not strain HB8. It actually does not match any known strain of *T. thermophilus*, but is most similar to strain HB27. This is not consistent with the 3D structure images found on the same page. In addition, there are labeling errors. Finally, because of the inherent nonlinearity of such a numbering scheme, labeling approximately only 10% of the nucleotides, and the inherent static nature of a secondary structure image file, many residues would be ambiguously numbered permanently. Working with primary sequence data, secondary structures, and tertiary structures is a painfully tedious and manual process, of double or triple checking and cross-referencing the divergent data sets stored in three different computer programs.

An unrelated problem is that when the x-ray structure was made, the researchers chose to put all the Mg ions for the whole 70S ribosome, in a single chain, in the small subunit PDB file. They were also numbered sequentially starting from 1. Most structural analysis programs would not be compatible with this organization scheme. Normally, Mg ions are stored in the chain of the rRNA or rProtein that they are closet to. Preferably, they would be numbered starting at a number higher than the complete rRNA or rProtein sequence would be. This guideline has been found to be violated in another structure. It is common for the N and/or C terminal ends of a protein to not be resolved in an X-ray

structure. If the C-terminal end is missing, but Mg ions are included in the same chain and numbered purely sequentially, then the Mg ion would share the same residue number with the unresolved amino acid in the c-terminal end. This causes errors and confusion.

RiboZones solves these problems. A fixed tertiary structure is provided along with a fixed secondary structure. Interactive secondary structures in the form of RiboVision ensures that the exact residue number of each nucleotide can be determined, along with the natural numbering. A complete mapping of the natural numbering system to the *E. coli* numbering system, including for unresolved nucleotides is provided in easy to use formats. In total, the time needed for making secondary structure based figures for *T. thermophilus* has been reduced by one to two orders of magnitude depending on the complexity of the figure.

3.3 Sequence Alignment

The final version of the RiboZones alignment is based on an alignment made by SINA. [62] Some problematic sequences were added to the SINA alignment using MAFFT's add sequence feature. [63] The alignment was manually checked by hand, and validated against secondary and tertiary structures. Further details are described in chapter 4.2.

3.4 Engineering and Reverse Engineering of Secondary Structures

3.4.1 Reverse engineering secondary structures

When a new ribosome tertiary structure is published, the authors usually include an image of the secondary structure. Ideally, they would provide the secondary structure in the original XRNA or VARNA format. It would also be good to provide an SVG file

and CT file. These files would make it easy to process the structure into RiboZones format, and make it easy for anyone to revise the structure or draw on the structure. Unfortunately this doesn't happen and requests for these files go unanswered. Therefore, the original secondary structure files must be reverse engineered. A secondary structure file consists of three parts, 1) A sequence of letters, 2) X and Y positions for each letter, 3) the helical base pairing pattern.

The sequence is easily attainable from either the tertiary or the primary structure, or processed out as part of the reverse engineering procedure. The helical base pairing pattern can be approximated from covariation analysis, and adjusted from the provided secondary structure image. Alternatively, in the case of a tertiary structure being available, various RNA structure programs can calculate the pseudoknot-free secondary structure of the RNA molecule. Further adjustments can be made by hand.

Obtaining the X and Y positions of the letters in the secondary structure is the most difficult part. To do this, the secondary structure must be provided in a vector graphics format. Fortunately, it is common to find secondary structures in the form of PDF's, with embedded vector format graphics embedded. The first step is to edit the PDF file, removing extraneous images, and get each letter separated into its own text box. This process involves creative use of Adobe Acrobat, Adobe Illustrator, and text editors. Saving the image in SVG Tiny format results in a clean image.

The image can optionally be scaled and positioned to make it more similar to the other RiboZones images. Font type, color, size can be adjusted as desired. Most likely the secondary structure will need to be relabeled, but this can be delayed until after further processing.

At this point, it is imperative to check the ordering of the letters in the secondary structure. Depending on where the original secondary structure file came from, the letters might be out of order. They need to be in 5' to 3' order, or 3' to 5' order. To solve this problem, RiboZones has a utility function to aid in the "descrambling" process. Fortunately, the complexity of descrambling only depends on the layout of the secondary structure, not on the degree of scrambling.

RiboLab can process the descrambled SVG file and aid in rescaling and positioning. Combined with the base pairing list, an XRNA file can be simulated. From there, the secondary can be edited using XRNA or VARNA. It is recommended to add the partially processed secondary structure to RiboVision, as that aids in further study.

3.4.2 Templating secondary structures

RiboZones primarily uses templated secondary structures. The sequence alignment is used to template a new secondary structure onto a previous one. A lot of manual adjusting needs to be done in XRNA. The advantage is that the new structure looks as similar to the previous structures as possible. This is especially important because RiboZones uses a new style of secondary structures.

3.4.3 New style secondary structures

During development of RiboZones, it was discovered that Helix 26a for the LSU was not properly labeled nor represented. The domain architecture did not make sense. The two piece layout of the LSU was inconvenient and difficult to superimpose. A new secondary structure for the LSU was developed.[43] The SSU also needed adjustments. [44]

3.5 Data Analysis

Data analysis is primarily done through MATLAB functions, the collection of which we call RiboLab. Most functions are command line only. However, some GUI features have been developed. RiboLab needs more polishing before it is ready for public release. This is not a complete list of functions, but covers the major features.

3.5.1 Basic objects

RiboLab is object oriented, making use of MATLAB's ClassDef feature. The lowest level object is a PDBentry. MATLAB has built in functions to parse a PDB file into a structure. The PDBentry class adds some additional functionality and stores these structures.

Other objects can be created from the PDBentry. Most usefully, structures can be broken into chains and individual residues. Most of the time, we use FullAtomModel representations of structures, where the atomic coordinates of all atoms are kept. However, there is support for a more coarse-grained model of the nucleotides. An individual nucleotide can be reduced to a PseudoAtom. There is a lot of flexibility in how the PseudoAtoms work.

3.5.2 Secondary Structures

The secondary structure object is currently called Map2D. The object contains everything needed to draw a secondary structure: nucleotide name, letter, X and Y position, etc. Secondary structures can be parsed from CSV files, XRNA files, or SVG files. Secondary structure objects can be outputted to EPS files, which can be processed into Adobe Illustrator, or as CSV files which can processed into RiboVision.

Alternatively, secondary structures could be saved as XRNA or CT files. Internally, the Map2D object has many uses.

3.5.3 Onion Objects

Initially, the onion object was designed to reproduce the methodology of Hsiao et al.[64] It allows for dividing the ribosome into concentric spherical shells, and calculating various properties as a function of shell. We added several features. The most notable feature is the removal of the shell restriction. The ribosome can be divided up into arbitrary subsets.

3.5.4 MapContacts

MapContacts is an extremely useful part of RiboLab. MapContacts takes one or two structures, and makes a list of which residues contact which residues, based on interatomic distances. At the core of MapContacts is an efficient nearest neighbor algorithm.

MapContacts itself contains data filtering options. Inputs can be filtered by residue subset or by atom type filters. Outputs can be filtered based on number of interactions found at the nucleotide level. In addition, many higher-level RiboLab functions filter the inputs and/or outputs to MapContacts.

3.5.5 CADS

The most mature part of RiboLab is the CADS object, which stands for "Conservation Analysis Data Set". CADS has evolved into being the central object in RiboLab. CADS objects are designed to store, organize, and analyze, sequence level data, and 3D structure data. They contain a Map2D, a target structure for MapContacts, and

subset information, for dividing molecules into multiple parts. They contain the results of analysis. CADS objects are designed to contain all the information needed for an experiment. Many RiboLab functions are designed to take CADS objects as their primary input.

3.5.6 Sequence Entropy

Seq_entropy is a function for calculating the entropy of a given multiple sequence alignment. It has many options. It can calculate Shannon entropy, something we call mutation entropy, and blossom adjusted entropy. It can calculate entropy based on individual letters, or on classes of letters. It has several methods of dealing with gaps and ambiguous nucleotides. Several higher-level RiboLab functions make use of Seq_entropy in unique ways.

CoVarEntropy is a function for calculating base-pair adjusted entropy. It takes a BP (Base Pair) structure as input along with the alignment. It can calculate several different definitions of base-pair adjusted entropy.

3.5.7 RiboLab_RV

RiboLab_RV is one of the better developed GUI's designed to help process new datasets explicitly for use in RiboVision. It takes many input files, but most are optional. RiboLab_RV's main inputs are a PDB file and a secondary structure file in RiboVision format. Additionally, domain definition files, helix definition files, alignment files, or FR3D files can be provided. In addition to processing the provided files, it calculates onion models, protein contacts, and magnesium contacts.

RiboLab_RV outputs the main table of a RiboVision dataset, in CSV format, ready to upload to the MySQL database. It also outputs the interactions tables if those

options are selected. A future version will be expanded to help with processing the secondary structure file including labels, and making the other files needed to set up a new RiboVision dataset. It is hoped that by providing these utilities, other people will help process datasets into RiboVision.

3.6 Dissecting the Ribosome

3.6.1 Helicoids

Early attempts at an evolutionary model were going to be based on using the traditionally defined helices in the ribosome as evolutionary units, with a few additional helices added. We wanted the "single-stranded" rRNA to be assigned to a helix. Every nucleotide would be in exactly one helix. Since this construct is not technically a helix, it is slightly more than a helix, we named them helicoids. Another condition, was that each helicoid could only be part of one rRNA domain. These rules together with the discovery of Helix 26a, necessitated a complete redesigning of the secondary structure of rRNA and the domain architecture of the large subunit.[43]

The helicoids were a useful construct. They allowed convenient drafting of figures. They allowed interactions to be grouped by helices and domains. They facilitated comparisons between species. Helicoids are a good size for exploring ribosomal structure. They can be turned on and off independently in PyMOL. A few helicoids at a time is the size of rRNA that can realistically be visualized at once. Early versions of our evolutionary models were based on helicoids.

3.6.2 Ancestral Expansion Segments

As our understanding of the ribosome matured, it became clear that helicoids were not the ideal evolutionary or structural unit. A near structural unit called ancestral expansion segments (AES's) are better suited for evolutionary studies (**CHAPTER 6**). Therefore, AES's are also provided in the PyMOL script suite.

3.6.3 PyMOL scripts

We have developed a suite of PyMOL scripts to make structure exploration and figure generation more convenient. The core script is called Master.pml. This script loads up pre-aligned pdb files, sets up useful PyMOL settings, and sets up default coloring schemes and representations for the whole pdb objects. It also separates the component rRNA pieces into their own objects, such as 23S, 5S, 16S, p-site tRNA, etc. Most of the higher scripts will operate on these intermediate objects. Therefore, changing the name of the PDB file itself will have no effect on other scripts. Also, the other scripts generally won't have to keep track of which chain is which piece of rRNA.

The other low-level scripts break the ribosome down into smaller component pieces. There are scripts containing many different definitions of the domains. There are scripts defining the "Helicoids". There are scripts separating out ions, water molecules, and other ligands. There are scripts to separate out the rProteins. There are even scripts to help define the exit tunnel.

Intermediate level scripts form the base workspace. They run the low-level scripts and set up other conditions. Intermediate level scripts make switching species easier. Running all available scripts for each species upon every load of PyMOL would take too long and cause too many PyMOL objects. Therefore, the typical workspaces included just

the basic level objects of multiple species, and detailed objects of one species in particular. Additionally scripts could be ran from inside the workspace as needed.

Higher level scripts are experiment specific. They are created for setting up for a specific type of analysis, tell a specific story, or render a specific figure. Therefore, these will not be included in the public release of RiboZones, except a few as examples.

3.7 Data Visualization

We present a tool devoted to the ribosome. This web-based portal, called RiboVision, is intended for rapid analysis, retrieval, filtering, and display of a variety of data types, simultaneously on three levels: primary (1D), secondary (2D), and three-dimensional (3D) structure, from ribosomes of six different species. RiboVision allows mapping and display of new and pre-loaded data, swapping of data between species, and quick generation of publication-quality images on any level of structure. RiboVision can maintain and display multiple layers of information, with controllable transparency. RiboVision allows users to import data from simple CSV format files and to map it directly onto all levels of structure. RiboVision has features in rough analogy with web-based map services capable of seamlessly switching the type of data displayed and the resolution or magnification of the display. RiboVision is available at http://apollo.chemistry.gatech.edu/RiboVision. We invite data for deposition, which would be made publically available.

This portal addresses the challenge of making accessible and integrating information on this key component of all biological systems. The basic features of RiboVision, such as selection of nucleotides, mapping the preloaded data onto 2D and 3D structures, and generating the figures, are intuitive. However, advanced features, such as

mapping multiple custom data, manual coloring of nucleotides and interactions, asynchronous display of 2D and 3D structures, require reference to documentation available at http://apollo.chemistry.gatech.edu/RiboVision/Documentation.

3.7.1 Features

RiboVision contains preloaded ribosomal structures along with preprocessed information related to these structures. The left menu (Figure 1a) controls data display, data import, and output. The right side toolbar controls data-visualization and website options. A detailed description of the toolbar is given in the website manual (which opens in a separate browser tab upon clicking on the RiboVision logo).

3.7.1.1 User Interface

RiboVision's interface (**Figure 3.1**) is based on modern web technology (jQueryUI) and design principles. Simple features should be easy to discover, with more advanced functionality available, and described in detail in the online documentation.

3.7.1.2 Included Data

RiboVision includes datasets for each of the species described earlier (**CHAPTER 3.21** and **CHAPTER 3.2.2**).

Information in the data menus is populated with pre-computed data sets. The default database currently contains nucleotide attributes that fall into three categories: Nucleotide Data, Phylogeny Data, and Protein Contacts.

Nucleotide Data allows glyph coloring and selection by (i) nucleotide number, (ii) helix or domain, (iii) radial distance from a geometric center [known as the onion

partition],[58] (iv) crystallographic B-factor, and/or (v) proximity to a magnesium ion (within 2,4 Å, 2.6 Å, or 6.0 Å).

Protein Contacts allows glyph coloring and selection by molecular interactions with any desired subset of rProteins.

Phylogeny Data allows glyph coloring and selection by conservation statistics obtained from a preloaded multiple sequence alignment of rRNA sequences using the RiboZones alignment. Phylogeny data are represented by pre-computed Shannon entropies (defined in a range 0-2).

Inter-Nucleotide Contacts allows visualization of various types of molecular interactions between nucleotides. The basic types of interactions *include base-pairing, base-stacking, base-phosphate, and base-sugar*. Each of these interactions types is further sub-divided using the data and nomenclatures of FR3D portal.[65] This menu also contains data related to RNA-RNA interactions mediated by proteins or magnesium ions, organized in *protein interactions* and *Magnesium2.4A* options. The interaction data between rRNA and rProteins can be limited by the user to any selected subset of proteins. Secondary structures do not explicitly reflect tertiary interactions, which are inferred from 3D structures. Tertiary interactions are mapped and visualized by lines on 2D structures according to their Inter-Nucleotide Contact attribute.

3.7.1.3 ResidueTip

ResidueTip is activated by hovering the mouse directly over a nucleotide glyph. Various information about that nucleotide is raised in the tip. The name of the selected nucleotide is given at the top, followed by the name of the consensus nucleotide. Next, the “Selected Data” is displayed. Selected Data is associated with that residue in the

currently selected layer. The Shannon entropy and a nucleotide frequency bar chart are also included. This information is very useful to have at hand while exploring other mapped information. Holding down the Alt key activates the ResidueTip feature for interactions between glyphs. This function shows ResidueTips for both nucleotides, and elaborates on the type of interaction (**Figure 3.1b**).

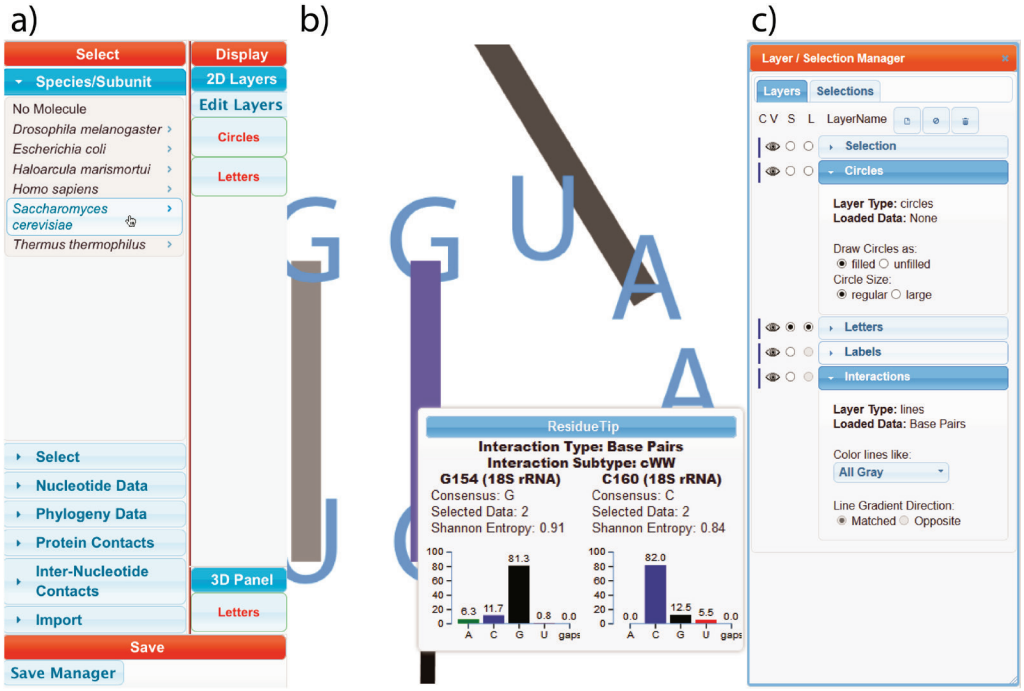

Figure 3.1. Main menu. The **Species/Subunit** menu offers selection of the LSU or SSU from six species. **Nucleotide Selection** provides options for selection and display of specific fragments of rRNA. **Nucleotide Data** contains nucleotide-specific data from previous structural analyses of ribosomes. **Phylogeny Data** contains the Shannon entropies of each nucleotide. **Protein Contacts** allows users to map interactions between rProteins and rRNA. **Inter-Nucleotide Contacts** allows the users to display interactions between nucleotides by type. **Import** allows users to upload data for mapping onto each level of ribosomal structure. **Display** contains layer objects to which data can be loaded by dragging and dropping from the Data menus. **Save** allows export of figures, along with additional saving and exporting options. a) **Species and subunit selection**. b) **ResidueTip**. Hovering the mouse over data mapped on the 2D structure produces a ResidueTip, a pop up box containing nucleotide-specific data. Hovering over an interaction line with the Alt-key gives a ResidueTip with data on both nucleotides. c) **Layer/Selection Manager**. The Circles layer and the Interactions layer options panels are opened here, revealing advanced functionality.

3.7.1.4 Layer/Selection Manager

Multiple layers of data, of various types, can be simultaneously displayed and independently manipulated. Basic manipulation of data is implemented in the **Main** menu (Select and Display panels). The advanced display and output are controlled by the **Layer Manager** (**Figure 3.1c**) located in the toolbar. Sequence-related data is projected into the "Selected" Layer, controlled by the (S column of radio buttons).

Each layer has a type that determines how it is mapped onto the 2D structure. Data can be used to color a nucleotide letter, to color a nucleotide circle, or to draw lines between nucleotide glyphs. The layer type can be seen by clicking and expanding the properties of a layer. The default layers are displayed in the Main menu. In the "Letters" layer, data are represented by colors of the nucleotide letters. In the "Circles" layer, data are represented by the colors of nucleotide circles. The users can create additional layers using the **Layer Manager**. User-added layers cannot be of type letter. The layers will also appear in the Display panel of the **Main** menu. The ordering of the layers is controlled by dragging layer objects in the Display panel of the **Main** Menu or in the **Layer Manager**.

The color on the 3D structures can be made synchronous with any 2D layer by setting the "Linked" property (L column of radio buttons, Figure 1c) for that layer. A layer can be temporarily turned off and on by clicking on the "V" property (the eye icon), or permanently deleted from the menu by clicking the Trash icon.

3.7.1.5 Coordinated Nucleotide Selection and Coloring

Selection of nucleotides is coordinated across the 2D and 3D panels. The user can add nucleotides in the 2D panel view to the active selection group. Nucleotides can be

selected individually or by using a click-and-drag selection box. A selected nucleotide is highlighted by a maroon circle. Multiple selection groups can be created and named using Selections tab of the Layers/Selection Manager. The user can re-activate a given group and add/remove nucleotide from the active selection group. In the 3D view, the selected nucleotides can be highlighted, or the user can choose to hide the non-selected nucleotides. Custom selections can be saved in a session in Save Manager.

3.7.1.6 Importing User Data

The **Import** menu allows users to import their own data, such as conservation frequencies, SHAPE reactivity, foot-printing data, etc.). User data can be quickly and accurately mapped onto 1D, 2D, and 3D levels of structure. Data are imported as CSV files. User data does not go to the server; it is read and stored locally on the client. Pre-generated CSV input data templates are provided on the website. Differences in the numbering schemes between species require specific templates for each species, which are provided (see the detailed tutorial in the RiboVision Manual for additional information.) Users can also import data from the CSV files generated by the Save Manager. These files may contain not only the species sequences but also nucleotide and interaction data.

In the CSV template, the resNum column specifies the nucleotides in a format MoleculeName:NucleotideNumber(s) [*e.g.* 5S:35, 16S:(49-578), or 23S:271M]. Following convention, in some species NucleotideNumbers can consist of both digits and letter characters. The second column can have the heading DataCol, which contains numerical values corresponding to a desired property, or the heading ColorCol, with either hexadecimal color codes or supported color names (1700+ colors, see manual).

When ColorCol is not provided, numerical information from DataCol will be mapped to a rainbow scheme, mapping the minimum value to blue, and the maximum value to red. Alternatively, the user can supply their own color scheme.

3.7.1.7 Saving Figures and Work

A number of options are incorporated to allow facile production of publication quality figures. Using the **Figures** tab in the **Save Manager,** images in the 1D, 2D, and 3D panels can be individually exported in a number of file formats. Specifically, images from the 1D panel can be exported in SVG format. Images from the 2D panel can be exported as PDF, SVG, JPG, or PNG. For the vector formats (SVG and PDF), either the currently visible layer or all layers can be outputted. 3D images from the Jmol applet can be exported as JPG images, or can be exported to a PyMOL script. The program exports a ZIP file that contains the PyMOL script with a description of the current state and the necessary PDB files.

The **Sequences & Data** tab allows easy export of the actual data used in figure generation. The whole sequence is exported, along with subsets of sequences stored in the user selections. Additionally, any data used to create figures is also exported as a table.

Users can save their current work to disk for later retrieval using the **Save/Restore Manager** tool from the Tool Bar or by clicking the **Session** tab inside the **Save Manager**. All layers and selections can be saved. The default save location is the local browser LocalStorage cache. Restoring will reset the display including layers, selections; loaded data types, *etc.* but requires re-loading the user input data. This feature ensures user data privacy. Alternatively, users can save their work to a text file. This

method allows users to restore the work session without an internet connection, or to move between browsers and computers.

3.7.2 Programming Details

RiboVision is a cross-platform webapp that integrates several advanced web solutions. Processing of data is done with jQuery, jQuery plugins, and JavaScript. The interface is based on jQuery UI. The 1D drawing is done in SVG with d3 library. The 2D structure drawing is done using the Canvas element. The 3D drawing is handled by the third party Java applet, Jmol. Our data are stored in a MySQL database and retrieved by a PHP server. The PHP server is also used to export file formatting and writing. Images are converted by ImageMagick. RiboVision functions, with certain limitations, on devices such as phones and tablets. RiboVision is under active development and still needs both client and server optimizations. The code is available on https://github.com/RiboZones/RiboVision.

3.7.3 Examples

Here we demonstrate the basic functionality of RiboVision using examples. We highlight the major functions of RiboVision rather than describe the actual data used for the examples. A fully detailed description of features, as well as explanations of the methods used to generate preloaded data sets, is contained in the manual (http://apollo.chemistry.gatech.edu/RiboVision/Documentation)

3.7.3.1 Example 1. Mapping 1D data onto 2D and 3D structures

RiboVision allows users to map various data simultaneously onto all three levels of structure (1D, 2D, and 3D). In example 1 (shown in **Figure 3.2**), we assign a color to

each nucleotide glyph of the *E. coli* 23S rRNA. The color corresponds to the Shannon entropy precomputed for a set of species sampled over all phylogeny. This example, which in practice takes two steps and less than 30 seconds to accomplish, shows how to visualize, at nucleotide resolution, the degree of conservation of rRNA. The input data (as well as the data used in Example 2) are a feature of RiboVision. The data contains the nucleotide number and a conservation score describing variability over a multiple sequence alignment using 133 representative species that sparsely represent the Woese tree of life. The 1D display (**Figure 3.2a**) shows the phylogenetic Shannon entropy of each nucleotide, illustrating that rRNA nucleotides from 1400 to 1600 of the 23S rRNA of *E. coli* exhibit high variability. The 2D display (**Figure 3.2b**) shows that the central loop of Domain V and other non-duplex rRNA are generally conserved (dark blue); while double-stranded nucleotides are variable (red, paired nucleotides co-vary). The 3D panel (**Figure 3.2c**) indicates that the central core of the LSU is highly conserved (blue) and the surface regions are variable (red).

To generate Figure 2, select LSU rRNA of *E. coli* from the **Species/Subunit** section of the Main Menu (**Figure 3.1a**). Drag Shannon Entropy from **Phylogeny Data** and drop it to the Circles layer object in the Display panel of the Main Menu. In a similar fashion, essentially any type of data can be mapped simultaneously onto the 1D, 2D, and 3D structures. For example, RiboVision contains crystallographic data, in the form of B factors, which can be mapped in the same way. The pre-loaded database in RiboVision will grow over time.

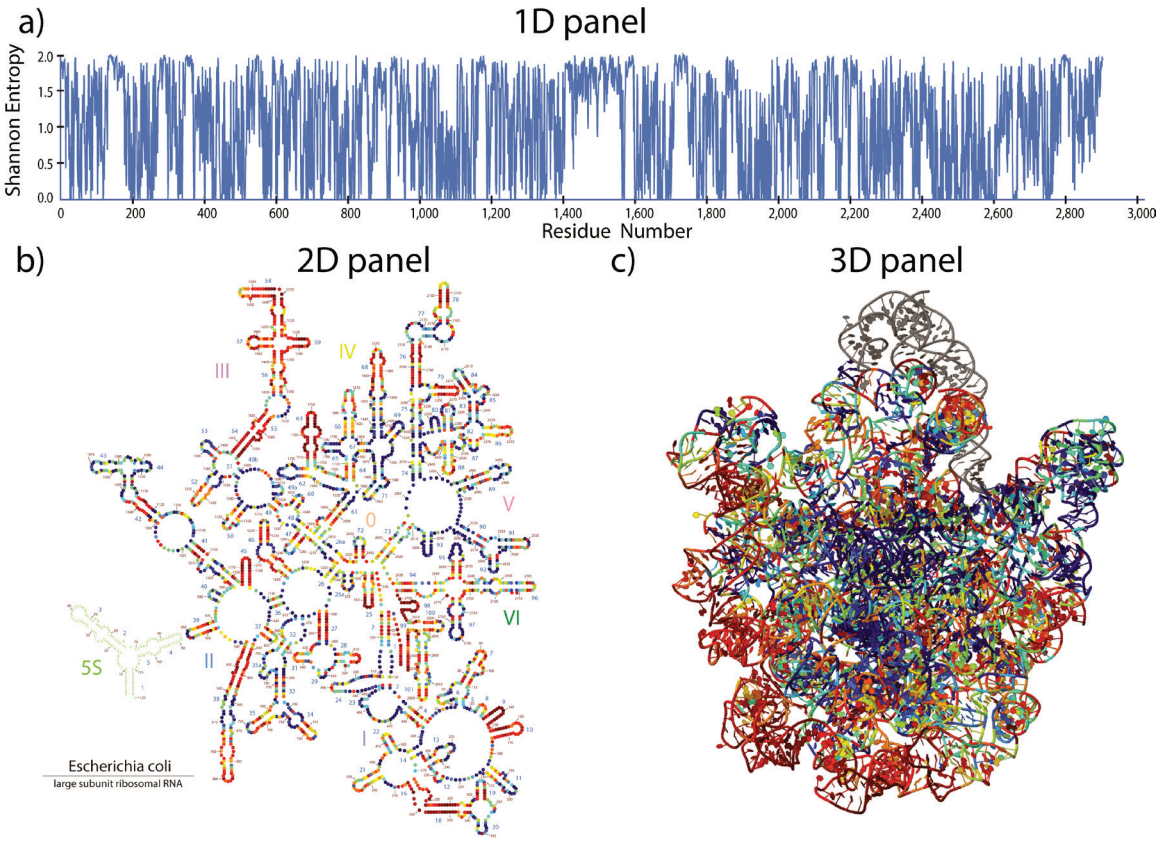

Figure 3.2. Mapping of Shannon entropies simultaneously onto 1D, 2D, and 3D structures of the *E. coli* 23S rRNA. Each nucleotide is assigned a color based on its Shannon entropy (the lowest values are blue; the highest values are red). The pre-computed Shannon entropies are plotted by nucleotide number in the a) 1D Panel, and mapped onto b) the 2D structure, and c) the 3D structure. The 23S rRNA nucleotides are numbered from 1 to 2904 and the 5S rRNA are numbered from 2905 to 3024 (Shannon entropies are not shown for 5S rRNA). Virtually any quantitative, nucleotide resolution data can be quickly mapped in this way.

3.7.3.2 Example 2. Mapping and visualizing the protein interactions in combination with other structural data

In this example, we illustrate a simultaneous mapping of the rRNA domain structure (from **Nucleotide Data**) and rRNA-rProtein molecular interactions (from **Protein Contacts)** onto the SSU rRNA of *T. thermophilus*. **Figure 3.3a** depicts the 2D structure of the 16S rRNA of *T. thermophilus* with nucleotide letters colored by domain (Domain 5' is light blue, Domain C is brown, Domain 3'M is pink, and Domain 3'm is green). The same 2D structure contains the information about the molecular interactions of rProteins with rRNA marked with colored circles. Nucleotides within 3.4 Å an rProtein are enclosed by circles that are color-coded by protein (S1, S2, etc.).

To generate Figure 3, select SSU rRNA of *T. thermophilus* from the **Species/Subunit** section of the Main Menu. Drag Domains from **Nucleotide Data** and drop it on the "Letters" layer object of the Display panel of the Main Menu. This action colors rRNA nucleotide letters by domain. Then select all proteins from the **Protein Contacts** menu, drag Protein Contacts object to the Circles Layer of the Display panel of the Main Menu. The name of the specific protein(s) interacting with a particular nucleotide can be seen in the **ResidueTip** window, which will automatically pop up upon hovering the cursor on top of a nucleotide glyph in the 2D structure.

Any given region of the ribosome can be represented in isolation and at high resolution. **Figure 3.3a** shows isolated Helix 23, and indicates that it interacts with rProtein S11 (light green circles). **Figure 3.3b** contains a selected region of the 2D structure of Helix 23. The nucleotides are selected using tools in the **Select** menu and marked by maroon circles. Nucleotides of Helix 23 that are in contact with rProtein S11

are highlighted by green circles. The same selected fragment, along with rProtein S11, appears on the 3D structure (**Figure 3.3c**).

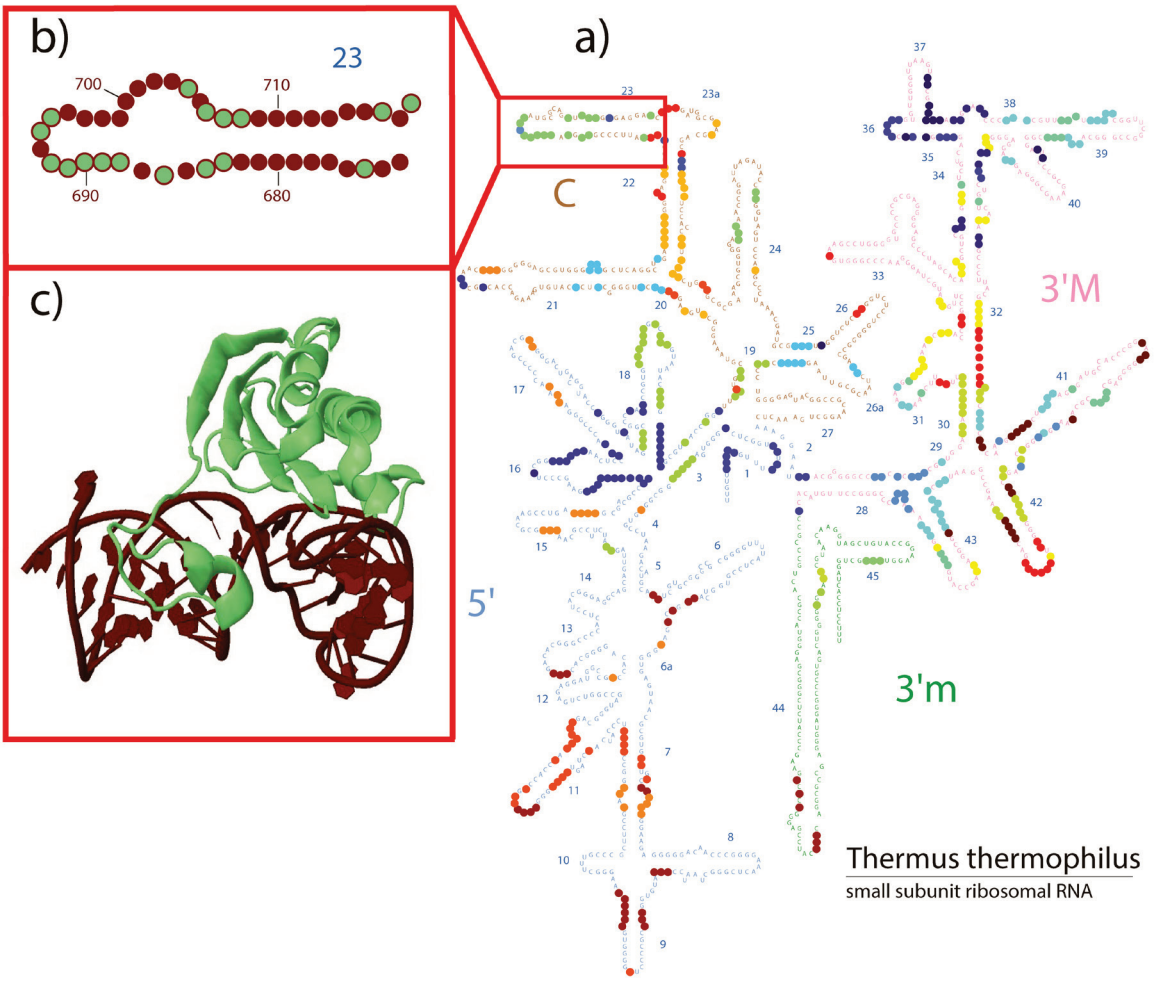

Figure 3.3 Visualizing interactions between ribosomal proteins of the small subunit of *T. thermophilus* and the 16S rRNA using RiboVision. a)The nucleotides in the 2D structure of 16S rRNA are colored by Domain (Domain 5' is light blue, Domain C is brown, Domain 3'M is pink, and Domain 3'm is green), while nucleotides contacting ribosomal proteins are overlaid with colored circles; each protein is assigned a distinct color. Interactions of rProtein S11 (green) with Helix 23 (maroon) of the SSU rRNA b) projected onto 2D structure and c) shown as cartoon representation of 3D structure.

3.7.3.3 Example 3. Visualizing user data and nucleotide interactions

RiboVision allows one to not only import and display data associated with single nucleotides but also to visualize pairwise nucleotide interactions, drawing lines between them on the 2D structure. The interactions can be direct (*e.g.* base pairs or base stacking) or mediated (*e.g.* by proteins or magnesium ions).

In Example 3, we demonstrate visualization of base pairing interactions in the 18S rRNA of *S. cerevisiae*. The interactions are indicated by gray lines and assign a color to each nucleotide according to type: A-green, G-black, U-red, and C-blue (**Figure 3.4**). The color definition is not preloaded to RiboVision but supplied by an external file (File Example3.csv is given in the Supplementary Data).

To achieve the display in **Figure 3.4**, select the SSU rRNA of *S. cerevisiae* from the **Species/Subunit** section of the Main Menu. First, import the Example3.csv containing assigned colors of each residue according to its identity [A, G, U, C], (available in the Supplementary Data) by opening **Import** section of the main menu and selecting the file from a location on the local workstation. Then, drag the "User Data" object from the **Import** section of the main menu to the "Circles" layer object. Finally, choose the Base Pairs data set from the **Inter-Nucleotide Contacts** menu to visualize Base Pairs interactions between rRNA nucleotides. The base pairs interaction data are automatically loaded into the Interactions layer of the **Layer Manager** (**Figure 3.1c**). For this example, we further filtered the base pair interactions by choosing only a few sub-types. By default, the lines appear in gray. Colored lines are an advanced feature. Additional details about connected nucleotides and the interaction type is shown in the

ResidueTip window upon hovering a mouse over a line while holding the ALT key (the key is required to distinguish selection of nucleotides from selection of lines).

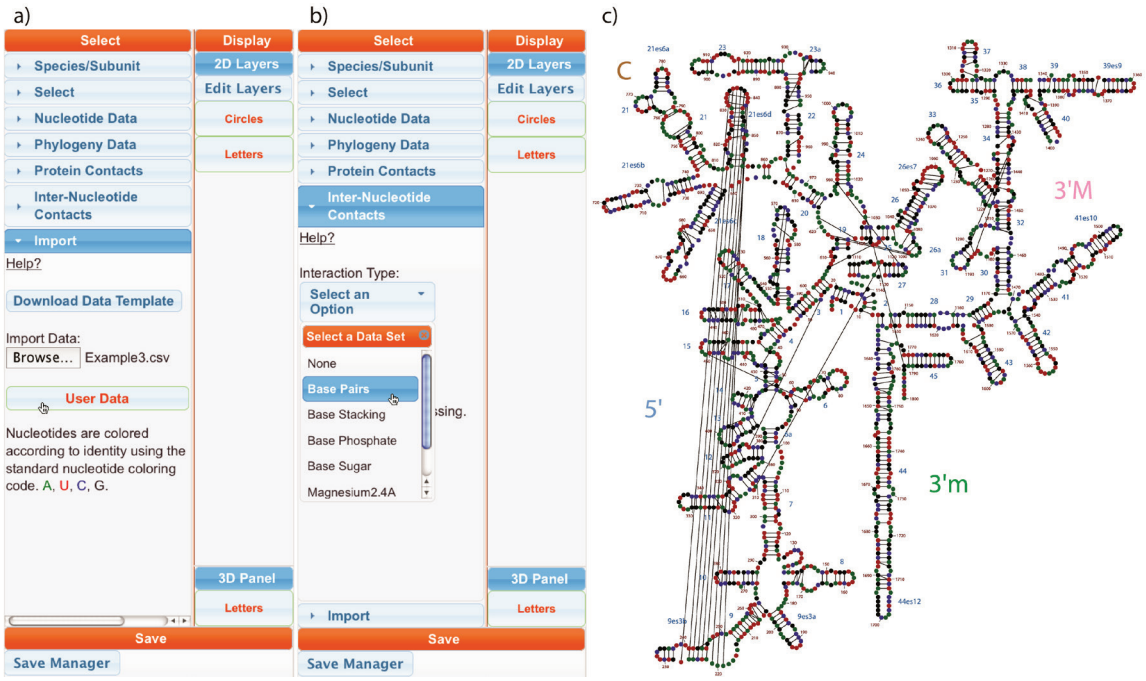

Figure 3.4. A subset of base pair interactions in *S. cerevisiae* 18S rRNA visualized along with imported user data. a) The user data file "Example3.csv" was loaded into RiboVision. A description of the file is visible under the "User Data" data object. b) Base Pairs were selected as the Interaction Type, from **Inter-Nucleotide Contacts**. Additional filtering was performed through selection of interaction subtypes. c) The resulting 2D display illustrates the standard nucleotide color code. Each nucleotide is assigned a color according to its identity. In addition, the gray lines connect nucleotides that form Watson-Crick / Watson-Crick interactions.

3.7.4 Conclusion and Future Perspectives

RiboVision was originally conceived as an in-house package for internal use in the Center for Ribosomal Origins and Evolution at Georgia Tech. We faced a bottleneck

in visualization and analysis of a variety of data-types on 1D, 2D, and 3D levels of structure, and in the rate of production of publication-quality figures. As the package matured, it was suggested that others might find it useful.

RiboVision has shown significant utility in our laboratories. This software enabled us to detect long-standing discrepancies between 2D and 3D structures of rRNAs. We have substantially revised the 2D structures of 23S/28S rRNAs[66] to reflect a double-stranded region (Helix 26a) formed between the left and right segments of the central loop in the traditional 2D structure. RiboVision enabled more modest revisions of 2D structures of 16S/18S rRNAs, providing accurate representation of the triple helical structure of the central pseudoknot and numerous non-canonical base pairs distributed throughout the 2D structure. We used RiboVision to create a public gallery of rRNA secondary structures, mapped with a variety of data (http://apollo.chemistry.gatech.edu/RibosomeGallery).

RiboVision is open source and is based on modern website technology. The version released here is focused primarily on manipulation and presentation at the 2D structural level with more limited options in 1D and 3D Panels. Visualization is limited to one subunit of one species at a time. The scope of the program will be extended in the future versions.

Ribosomes are not the only large assemblies with a high level of complexity and extensive available data. We believe that the framework here can be further generalized to protein structures, other macromolecular assemblies such as viruses and bacteriophages, and to organelles and cells.

CHAPTER 4

SEQUENCE ANALYSIS IS MORE POWERFUL WHEN INCORPORATING STRUCTURE

4.1 Introduction

The question, "how did the ribosome evolve?" is fundamentally a question about the structure and function of the ribosome. While comparative sequence analysis is a powerful bioinformatics tool, [38,67-69] it is not sufficient to answer this question alone. However, sequence analysis is very useful when combined with structural information. Thousands of ribosomal sequences are known awaiting thorough data mining. It is well known that the outcomes of analysis is heavily dependent on the MSA used for analysis,[70] so it is important to define what an MSA will be used for before choosing an algorithm.

There is an abundant amount of sequences available for rRNA, especially for the small subunit. The SILVA database (release 119) SSU reference set contains 1,583,868 sequences, and the SILVA database (release 119) LSU reference set contains 57,546 sequences.[36] There is a moderate number of secondary structures available. Based on the Comparative RNA Web Site and Project,[38] there are 663 SSU structures and only 85 LSU structures. Ribosomal 3D structures are much more difficult to obtain, requiring more resources. Only a handful of species have had any part of their ribosomes crystalized.

Multiple Sequence Alignments (MSAs) can be used to infer which parts of rRNA sequence are homologous between two or more species. The MSA also allows one to use known secondary and three-dimensional structures as templates to model rRNAs with unknown structures. RiboZones has used MSA's for both of these purposes as well as

other types of analysis. However, it is critical to use a structure based MSA to get accurate structural information out of the MSA.

Here, we describe how we made the RiboZones structure-based MSA. In general, secondary, and three-dimensional structures are more highly conserved than sequence. A structure-based MSA should reflect this independent observation. It should be possible to quantitate (score) how well an alignment matches a series of structures. The score could be used to iteratively adjust the MSA. The score is also a measure in the confidence of the ability to use the MSA to predict the structure of another species in the alignment.

4.1.1 Alignment algorithms

Our goal is to incorporate structural information in the alignment process. There are many different alignment algorithms[62,67,71-74] with various strengths and weaknesses, and specialized functions and types of analysis. We have devised an iterative process in which sequence alignments are used to infer structure information, then structural models are used to reevaluate the alignment.

Highly accurate alignments for rRNA require algorithms specialized for rRNA. Alignment algorithms which predict the secondary structure of the RNA, for example LocARNA[51], R-Coffee,[75], and FOLDALIGN[76] hold promise. However, these algorithms can only be as good as the structure prediction algorithms. Alignment algorithms, which may be well suited for smaller RNAs, fail on application to large ribosomal RNAs. They currently do not work well on the large, structurally variable, highly distorted RNA in the ribosome.

Specialized rRNA aligners all generally work by relying on a template alignment.[38,62] The templates are built manually by experts, and new sequences are

aligned into them through different algorithms. The quality of the result is dependent on the quality of the template alignment. A weakness of this approach is that although there are standard reference alignment test sets for proteins,[68] there is no such test set for LSU or SSU rRNAs.[77] There is no standard alignment nor is there a standard way to measure ultimate quality. Due to the variability of RNA structure, such an ideal standard would be difficult to discern, as sometimes it is difficult even with two sequences of known structure. In general however, stems, obvious bulges and insertions, and loops, should be visibly separable in the alignment.

4.1.2 Available Alignments

In practice, there are two de facto standard reference alignments: Gutell's CRW alignment,[38] and SILVA's seed alignment.[36] Either of these could be used as starting places for making one's own alignment suited for their purposes. For research primarily on animal sequences, there is also the Mallatt alignment.[78]

4.1.3 Status of the existing alignments

The CRW site provides a 3-domain alignment for both the SSU and LSU rRNA. However, it suffers from several issues. The sequences included are not well balanced, with several major phyla lacking representation. There are a many repeated sequences with minor variations, for example, there are 39 copies of *Escherichia coli*, and 5 copies of *Mus musculus*. Desirable species, including *Haloarcula marismortui* and *Homo sapiens* are not included. Approximately 10% of the species are incomplete. In addition, this alignment is not actively maintained. Many more sequences, especially those of eukaryotes became available since the CRW alignment was made. There is no official way for a user to add sequences to the alignment, although tools such as MAFFT could

be used. The CRWAlign program only offers a 16S bacterial template, which has limited use.

SILVA, a database of rRNA sequences along with seed and reference alignments, is actively maintained. Unfortunately, the seed alignment is not publicly available because it contains unpublished sequences. The seed alignment, like the CRW alignment, has been extensively adjusted, manually by experts. SILVA provides their own alignment algorithm, SINA,[62] which can add user sequences into their global reference alignment. The aligner attempts to recognize junk sequences and is aware of secondary structure, including in their scoring algorithms. SILVA's has some major difficulties with certain metazoan sequences. See **CHAPTER 4.5** for details.

The Mallat alignment specializes in metazoan sequences. It is a detailed structural and phylogenetic based alignment. It contains many incomplete sequences, because that is the state of the art for available metazoan sequences. Most metazoan rRNAs are either partially sequenced or their LSU rRNA genes are only partially annotated. The file format is useful for phylogeny analysis and for visual inspection. The file format is not compatible with RiboZones software at this time.

4.1.4 RiboZone philosophy

RiboZones is an actively maintained and developed database of both sequences and alignments. There is an extensive amount of meta-information available for each sequence. However, for RiboZones, a smaller set of fully sequenced species is desired. Achieving this, means reconstructing some eukaryotic sequences that have not been done before. The resulting set of species would be much smaller, but a good starting point for someone doing broad sequence analysis. The alignment should agree with known

structural alignments or ribosomes. RiboZones, in general, supports SILVA. This chapter describes the process of building the alignment, evaluating the alignment and the problems discovered with previous alignments.

4.2 Methodology

4.2.1 Sequence Criteria

Sequences must meet the following criteria. 1) Each sequence must be nearly complete, within 20 nucleotides of their estimated 3' and 5' ends. 2) Each sequence must be the sole representative for its species. 3) Each sequence must be from a fully sequenced organism, to facilitate the same species list potentially being used to make alignments of other genes, especially translation-related proteins. 4) Each sequence must sample the tree of life fairly, according to current data. 5) Each sequence must match what is in the 3D structures, when available. 6) Each sequence must have intervening sequences removed as much as possible.

4.2.2 Sequence Collection

In accordance with the sequence criteria, rRNA sequences for 133 organisms have been compiled for both the LSU and the SSU. Sequences were primarily taken from SILVA and NCBI databases. Some sequences took several days of effort each. Their sequences could only be reconstructed after searching multiple databases with multiple queries and tools, and piecing together data from several sources. As gene annotations were often not sufficient, contigs and chromosomal assemblies of DNA were examined to find missing pieces of rRNA. The exact 3' and 5' ends of rRNA had to be approximated based on observed patterns. For species with fragmented rRNA, all

fragments were combined into the correct order. Some sequences had extraneous sequence wrongly included in their annotations, such as ITS2 or other spaces. The Trypanosomes were especially troublesome. Trypanosomes have several extra legitimate expansions. However, they also have a fragmented domain VI, which must be added in the correct order. Surprisingly, chicken was also difficult. The assembly of chicken chromosomes do not include a single copy of a whole ribosomal subunit. Ribosomal RNA is often in repeat regions are not included properly in assembles of contigs.

We have developed the most complete and accurate MSA of SSU and LSU rRNA sequences available. The sample size, 133 is relatively small, but adding more sequences is relatively simple. MAFFT has been shown to work well using the RiboZones MSA as a template.

4.2.3 Initial alignment

Alignments were initially produced using SINA. Sequences with 5.8S/Domain I problems were removed from the SINA alignment. MAFFT was used to add these problematic sequences back into the SINA alignment. MAFFT aligned these sequences well, which looked just like SINA had aligned them correctly. Manual adjustments to the alignment were done to correct alignment problems and make the alignment better match the 3D structure.

4.2.4 Calculating predicted base pairs (base pair entropy)

In a perfect alignment, most base pairs in one species should be predicted in another species, at least for homologous rRNA. This means that if two columns in an alignment are base paired, in for example, *E. coli*, most likely it should be base paired in *S. cerevisiae*. Sometimes, this is not true, as structural changes sometimes happen

between taxa. The final alignment should reflect what is known to be true structurally. However, a poor alignment will falsely under predict base pairs.

A modified version of Shannon entropy[79] is used to determine how well predicted each base pair is. A set of known base pairs for a particular sequence/structure is needed; here *E. coli* is used. The known base pairs are any base pair marked as cWW by FR3D.[80] FR3D analysis was taken from the RNA 3D Hub database.[48] A base pair represents two columns in a MSA. Each position can have one of 5 values, A, G, C, U, or – (gap). Therefore, there are five squared or 25 possible combinations of two characters, dyads. This is a rough statistic, so to a first level approximation, a (likely) base pair can be defined as these combinations, AU, GU, AG, CG, CA, UU, in either order. For each known base pair, the frequency of each of the 25 types of dyads is calculated. The dyads are classified into two groups, either in the defined base pair list, or not. Frequencies are converted to entropy. The base pair entropies are visualized using RiboVision.[52] Positions with high entropy can be a true negative, a base pair has been lost amongst many species or a false negative, there should be a predicted base pair, but one or more domains are not aligned properly. If non-base pairing combinations come up, like CC, and A-, they are more likely to be caused by misalignment rather than legitimate structure changes.

4.2.5 Calculating Structural Divergence

Structural divergence (RMSD) will be calculated from superimposed 3D ribosome structures. Ribosome structures were obtained from The DARC site: a database of aligned ribosomal complexes.[47] *E. coli* LSU (3R8T), *E. coli* SSU (3R8O), *S. cerevisiae* LSU (3U5D), and *S. cerevisiae* SSU (3U5B) pdb files were obtained from DARC. LSU

structures were processed as is. SSU structures were realigned using PyMOL's cealign command.[81]

Structural divergence is the distance between the atoms in one structure and the corresponding atoms in another structure. The square root of the mean of the squares of the structural divergence yields the standard RMSD. To easier pair up atoms and calculate this statistic on a per nucleotide basis, each nucleotide in the rRNA (except the 5S) was reduced to a pseudoatom. The pseudoatom was calculated as the center of mass of the set of phosphate, sugar, and N1 or N9 atoms of each nucleotide. Next, the alignment was used to calculate which residues in one species corresponded to which residues in the other species. This list was further reduced to contain only corresponding residues which both exist in the PDB files, and hence have known pseudoatom positions. The distance between the pseudoatoms of these corresponding pairs is calculated, and called the structural divergence.

4.2.6 Calculating Gap Frequency

RiboZones entropy programs have the ability to overwrite the entropy for any position along the alignment where a threshold gap frequency has been exceeded. Manual manipulation of the overridden entropy data is used to produce figures. For example, it is possible to visualize just the positions with high gaps by not including positions with low gaps in the user data file. Alternatively, the values for the high gapped positions can be made artificially high or low, to force an obvious color differential with the rest of the entropy data. For the examples in this chapter, nucleotides were put into one of two categories, at least 20% gap characters, or less than 20% gap characters.

4.3 Building and evaluating a structure-based alignment

SILVA accurately aligns most helices with respect to each other, within a domain of life: bacteria, archaea, and eukarya. A few helices are more difficult and exhibit less accuracy and confidence. A few species, most notably those with less common expansion segments, require additional manual adjusting.

SILVA and the other alignment algorithms tested, produced inaccuracies in several helices, with respect to the relative alignment between the domains. Bacteria, Archaea, and Eukaryota have helices, bulges, and loops of varying lengths. It is difficult to notice and correct this based just on the alignment and normal entropy scores.

4.3.1 Using base pair entropy

RNA structural information can be used to score and guide the alignment. Our desired alignment best alignment maximizes the measurement of structural homology. Three parameters need to be optimized; 1) The number of predicted base pairs, 2) The RMSD between two superimposed structures, and 3) The number and distribution of gaps. RiboZones' reference species, *Escherichia coli*, and *Saccharomyces cerevisiae* have accurate secondary and tertiary structures available. These structures can be used to score and validate the alignment.

Mapping base pair entropy onto the secondary structure of the LSU (**Figure 4.1**) helps identify problem regions. Each problem region should be manually examined to determine the cause for the high variability. It should be determined if the high variability is a true signal, or if the alignment needs adjusting. Upon close examination of the data, Helices 9, 63, 68, 78, and 98 have variable length, causing high gap penalties in the variable length region. These helices become significantly shorter or nonexistent in a

large enough proportion (>5%) of species. Helices 10, 16, 18, 54, 55, 56, 58, 59, and 79 have alignment problems. These regions are structurally variable with helices of varying lengths, making it difficult to say for sure which parts everything has in common and which parts are the variable length regions.

Similarly, mapping base pair entropy onto the secondary structure of the SSU (**Figure 4.2**) helps identify problem regions. Each problem region should be manually examined to determine the problem. Some regions can be fixed with better alignment. Upon close examination of the data, Helices 6, 10, 17, and 44 have variable length. These helices become significantly shorter in a large enough proportion (>5%) of species. There are no major unalignable regions within the bacterial small subunit. Overall, the SSU is better behaved than the LSU.

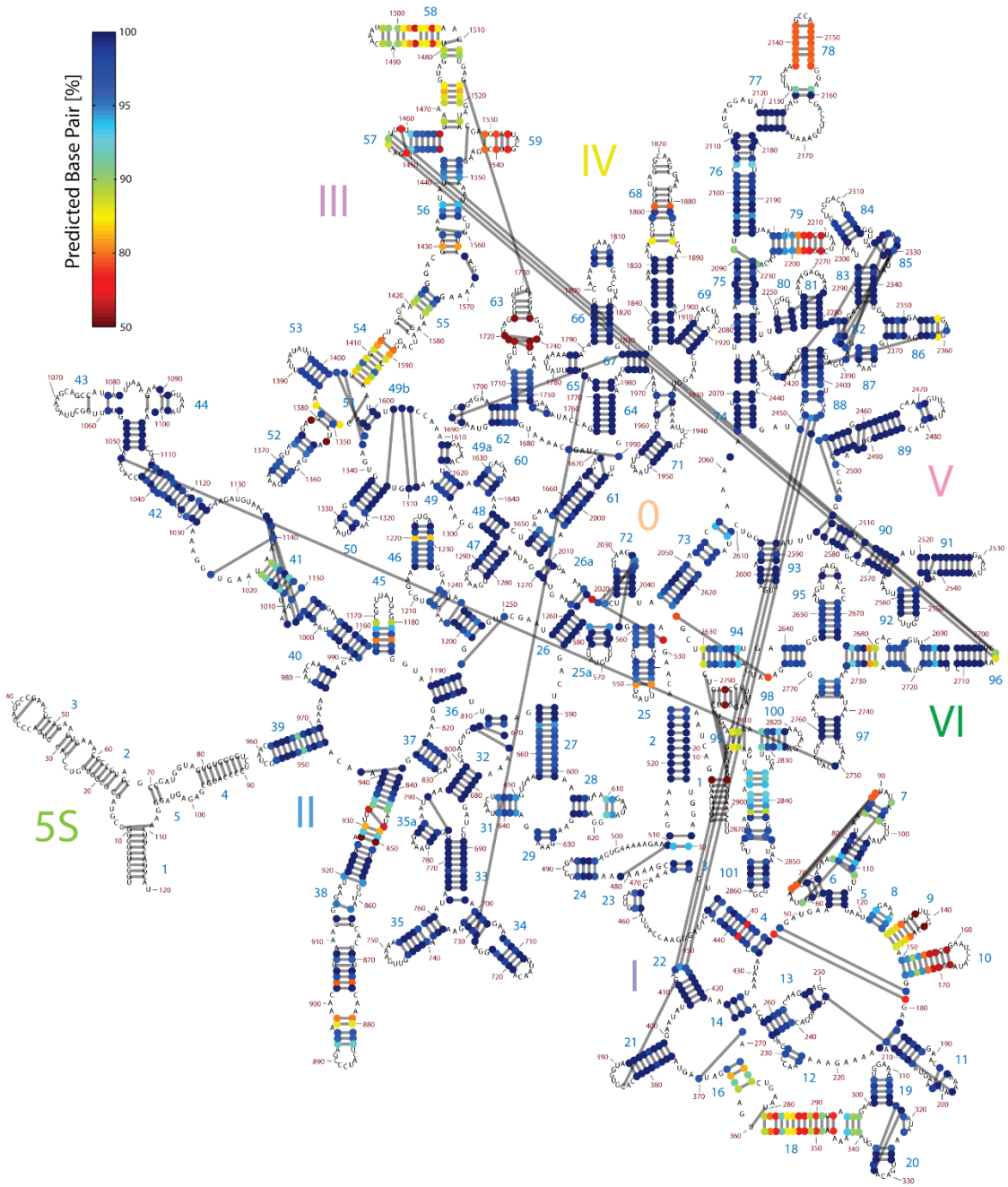

Figure 4.1. Base pair entropy (frequency) mapped onto the secondary structure of *E. coli* LSU rRNA. Red and orange base pairs are not predicted in a large percentage of the species. Often, these base pairs are only in bacteria or are in difficult to align regions. Helices 9, 63, 68, 78, and 98 can shrink or disappear. Helices 10, 16, 18, 54, 55, 56, 58, 59, and 79 are structurally variable and can contain expansion segments. They are more difficult to align properly. The black lines represent cWW base pairs as calculated by FR3D, taken from the RNA 3D Hub database.

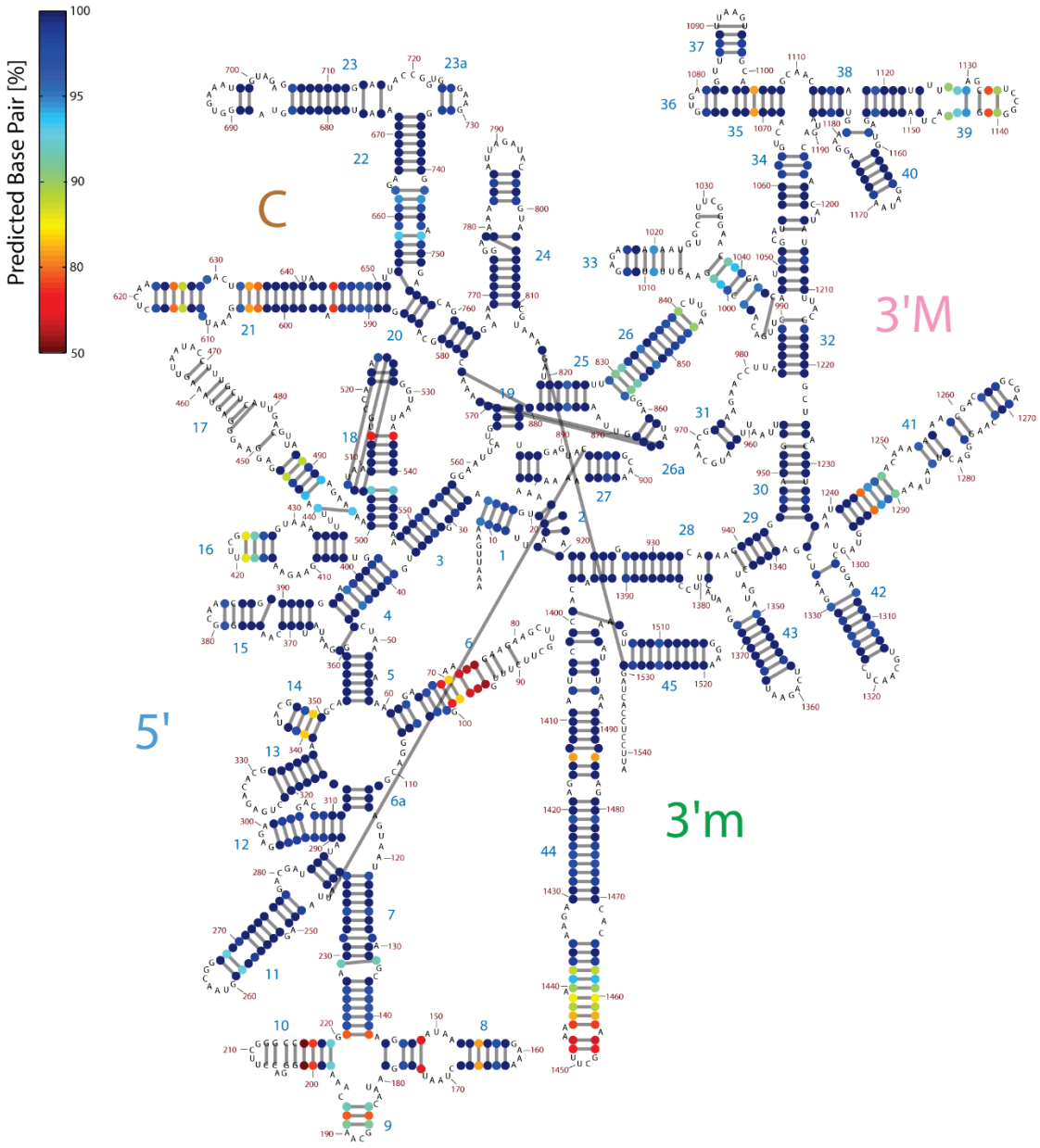

Figure 4.2. Base pair entropy (frequency) mapped onto the secondary structure of *E. coli* SSU rRNA. Red and orange base pairs are not predicted in a large percentage of the species. Often, these base pairs are only in bacteria or are in difficult to align regions. Helices 6, 10, 17, and 44 can shrink. Helices 9, 33, and 39 are structurally variable and can contain expansion segments. They are more difficult to align properly. The black lines represent cWW base pairs as calculated by FR3D, taken from the RNA 3D Hub database.

4.3.2 Using structural divergence

Structural divergence is a valuable statistic used to understand the structural pattern of rRNA. The traditional RMSD statistic is the square root of the mean of the structural divergences squared. Using the MSA as a guide, one can map the structural divergence between two known structures. Different alignments will exhibit different structural divergence patterns. In most cases, a lower structural divergence is indicative of a better alignment. Comparing structural divergence as a function of nucleotide is a useful way to compare two or more alignments. The regions where the alignments agree and disagree are clearly visible.

Structural divergence plots of the RiboZones alignment, the CRW alignment, and the straight SILVA alignment were used to iteratively refine the alignment. Regions where the alignments disagreed were manually examined. Individual sections of the ribosome rRNA were carefully superimposed by hand. The sequence alignment was manually adjusted, to better reflect the 3D structure.

It was decided that in some places, most notably the variable part (Helices 54-59) of Domain III, to not adjust the alignment. The structural similarity was not strong enough to align reliably the structures. In addition, there was not enough information to align properly all the other species in the alignment set in these regions. The overall structural divergence pattern and RMSD would not appreciably change, but the base pairing patterns would become less accurate.

Structural divergence was visualized by mapping it on the secondary structures of *E. coli* using RiboVision (**Figure 4.3** & **Figure 4.4**). Expanded helices in *S. cerevisiae* can be predicted based on the color of the tetraloops. There is no logical alignment for the

RNA loops of H10, H25, or H98. H43 and H44, the L11 binding domain, is light blue due to bending of L7/L12 stalk. This part of the ribosome is naturally mobile. H58 is green/yellow because this helix bends off into different directions between *E. coli* and *S. cerevisiae*.

4.3.3 Minimizing gaps

Gap frequency data helps minimize the gaps in an alignment. The alignment should have as few gaps as the data supports. SINA/SILVA sometimes introduces extra gaps. Some extra gaps are clearly visible through inspection of the alignment itself. Other suspected gaps require the 3D structure to aid in accurate adjustments. Mapping gap frequency onto the secondary structure is helpful. The gap frequency can be filtered to any level, meaning that above a certain frequency, for example 20%, all positions with 20% or more gaps will be marked as one color. The remaining nucleotides either can share a single color, or use a color spectrum. One strategy would be to do this process iteratively, making the gap filter lower each iteration. In this example (**Figure 4.5** and **Figure 4.6**), positions with less than 20% gaps have been marked as dark blue, and positions with 20% or more have been marked as dark red.

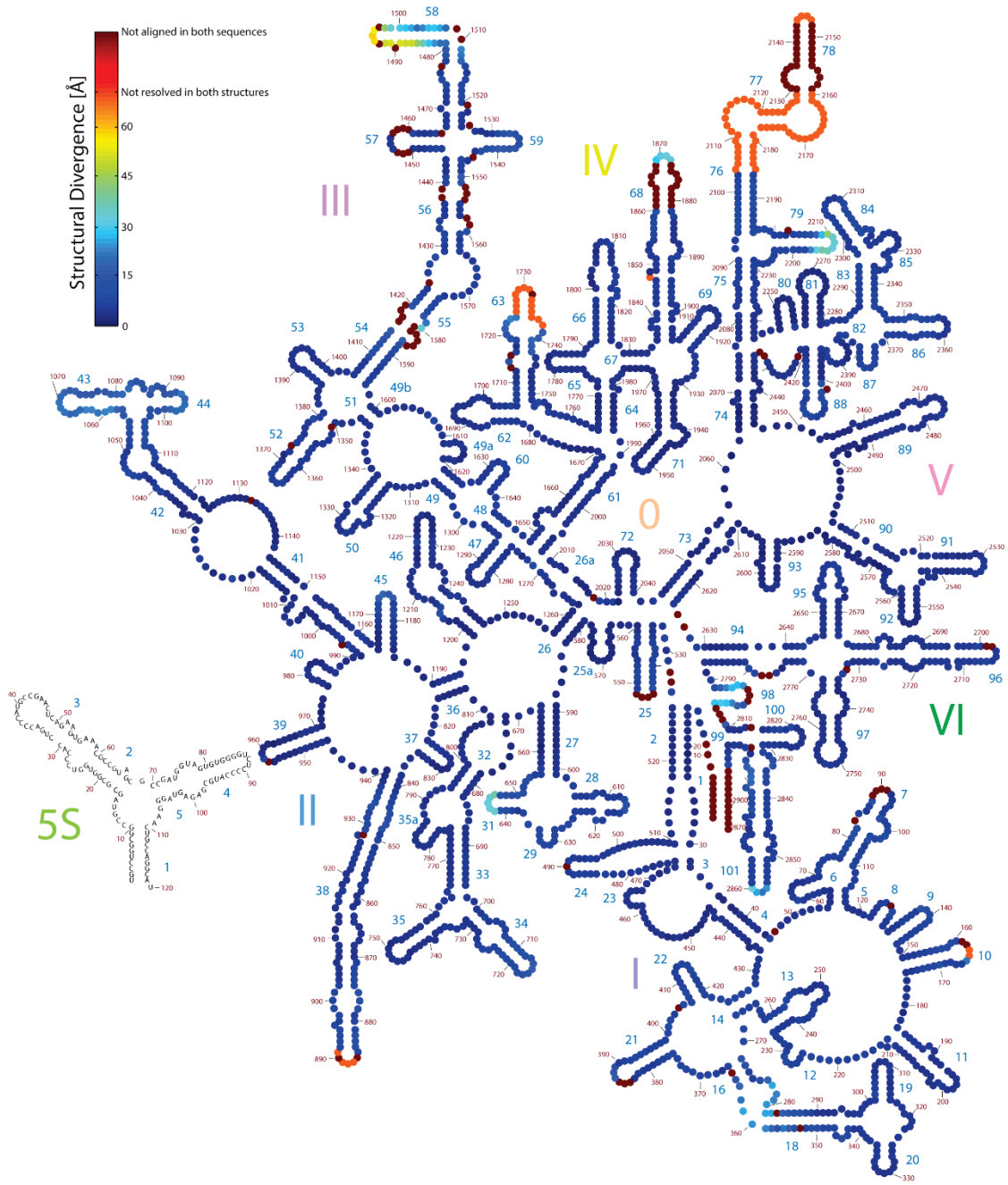

Figure 4.3. Structural divergence mapped onto the secondary structure of *E. coli* LSU rRNA. Dark blue means a good correspondence and superimposition. Dark orange means one or both nucleotides were not resolved in the PDB file. Dark red means no corresponding nucleotide in *S. cerevisiae* due to either alignment problems or legitimate deletions. Light blue / green tetraloops are a consequence of their respective helices growing longer in S. cerevisiae. There is no logical tetraloop for H10, H25, or H98. H33/34 is light blue due to bending of this region, possibly because of crystal packing effects. H58 is green/yellow because this helix bends off into different directions between *E. coli* and *S. cerevisiae*.

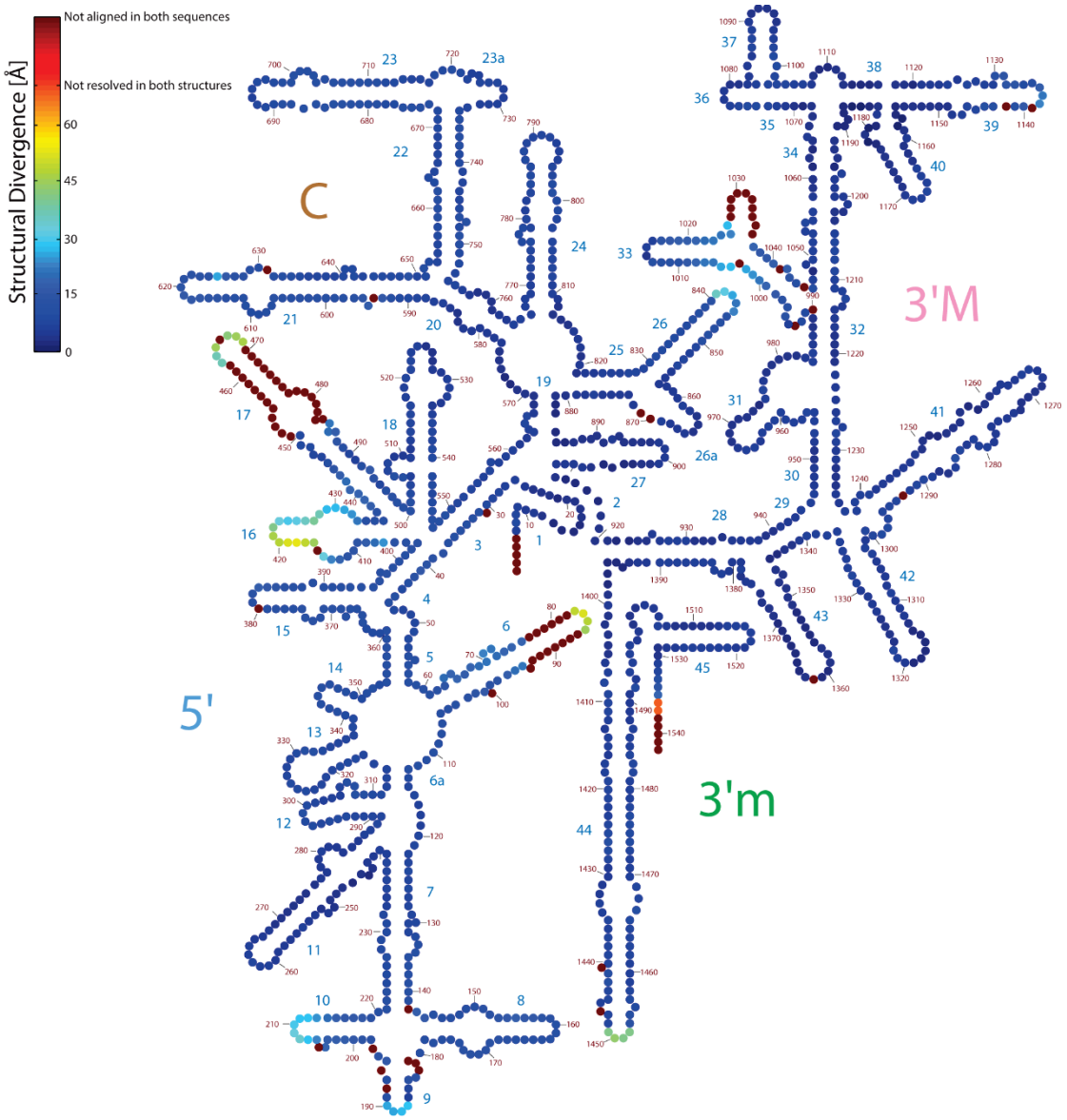

Figure 4.4. Structural divergence mapped onto the secondary structure of *E. coli* SSU rRNA. Dark blue means a good correspondence and superimposition. Dark orange means one or both nucleotides were not resolved in the PDB file. Dark red means no corresponding nucleotide in *S. cerevisiae* due to either alignment problems or legitimate deletions. Light blue / green tetraloops are a consequence of their respective helices growing longer in *S. cerevisiae*.

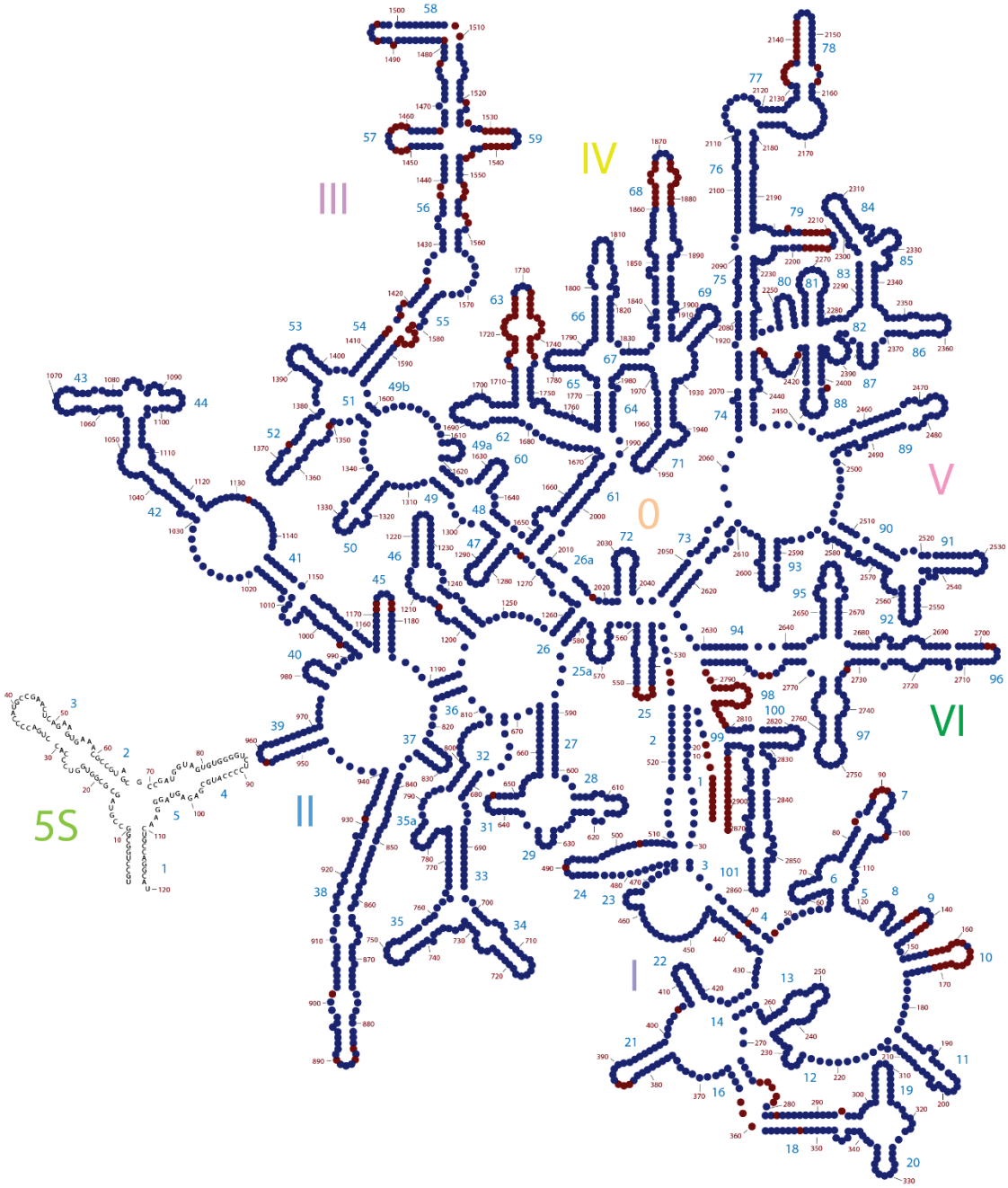

Figure 4.5. Gap frequency filter at 20% for *E. coli* LSU rRNA. Nucleotides whose position the alignment have less than 20% gaps are marked as dark blue, otherwise they are marked as dark red. The reason for all gaps should be documented. Ideally, there should not be any half-gapped base pairs, but there are a few here in difficult to align regions.

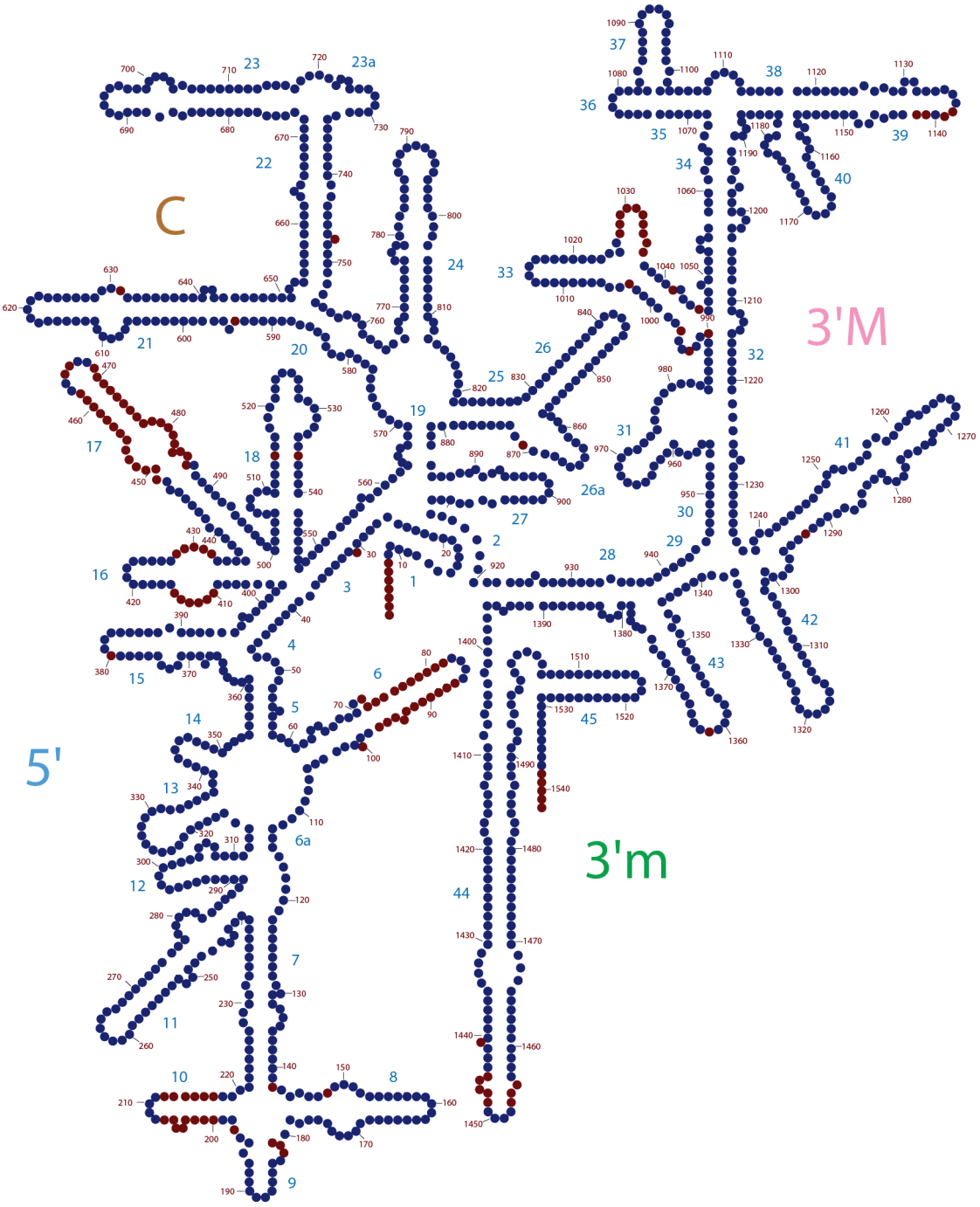

Figure 4.6. Gap frequency filter at 20% for *E. coli* SSU rRNA. Nucleotides whose position the alignment have less than 20% gaps are marked as dark blue, otherwise they are marked as dark red. The reason for all gaps should be documented. Ideally, there should not be any half-gapped base pairs, but there are a few here in difficult to align regions.

4.4 Visualizing a structure-based alignment

The alignment should be visualized considering base pair entropy. Using standard Shannon entropy on the rRNA is misleading and underestimates structural conservation (**Figure 4.7A&C**). Base pair adjusted entropy only covers a fraction of the rRNA, as there is a significant portion of single stranded rRNA (**Figure 4.1** and **Figure 4.2**). Combining both statistics allows for full coverage and a better estimate of structural conservation. The statistics can be combined by first using the regular Shannon entropy for all nucleotides, and then overwriting the entropies with base pair adjusted entropy for residues that have a predicted base pair frequency above a certain threshold percentage.

The alignment is visualized in **Figure 4.7B&D**. The threshold percentage was determined by considering the average entropy over the whole rRNA. The average combined entropy statistic for the whole LSU or SSU is graphed as a function of minimum predicted base pair percentage (**Figure 4.8**). A percentage of greater than 100% is equivalent to using only standard Shannon entropies and no base pair entropies. A percentage of 100% means that the base pair entropy will only be used if every single species has a predicted base pair, based on having a dyad sequence in the allowed classes, at that position. It is shown that using a percentage of 100% dramatically causes a drop in the average entropy as predicted. Every one of these base pairs would have an entropy of 0 as opposed to what the regular entropy would be, in the range of 0 to 2. However, using a threshold of 100% still overestimates the total entropy, due to no tolerance of exceptions. Sources of exceptions include sequencing error, alignment error, or a genuine mutation / loss of base pair. If these events are rare, that position should remain marked as a conserved base pair position. Based on the data and traditional scientific procedure, a

threshold of 95% was chosen. Lowering the threshold further would not cause a significant change in entropy worth the decrease in data considered.

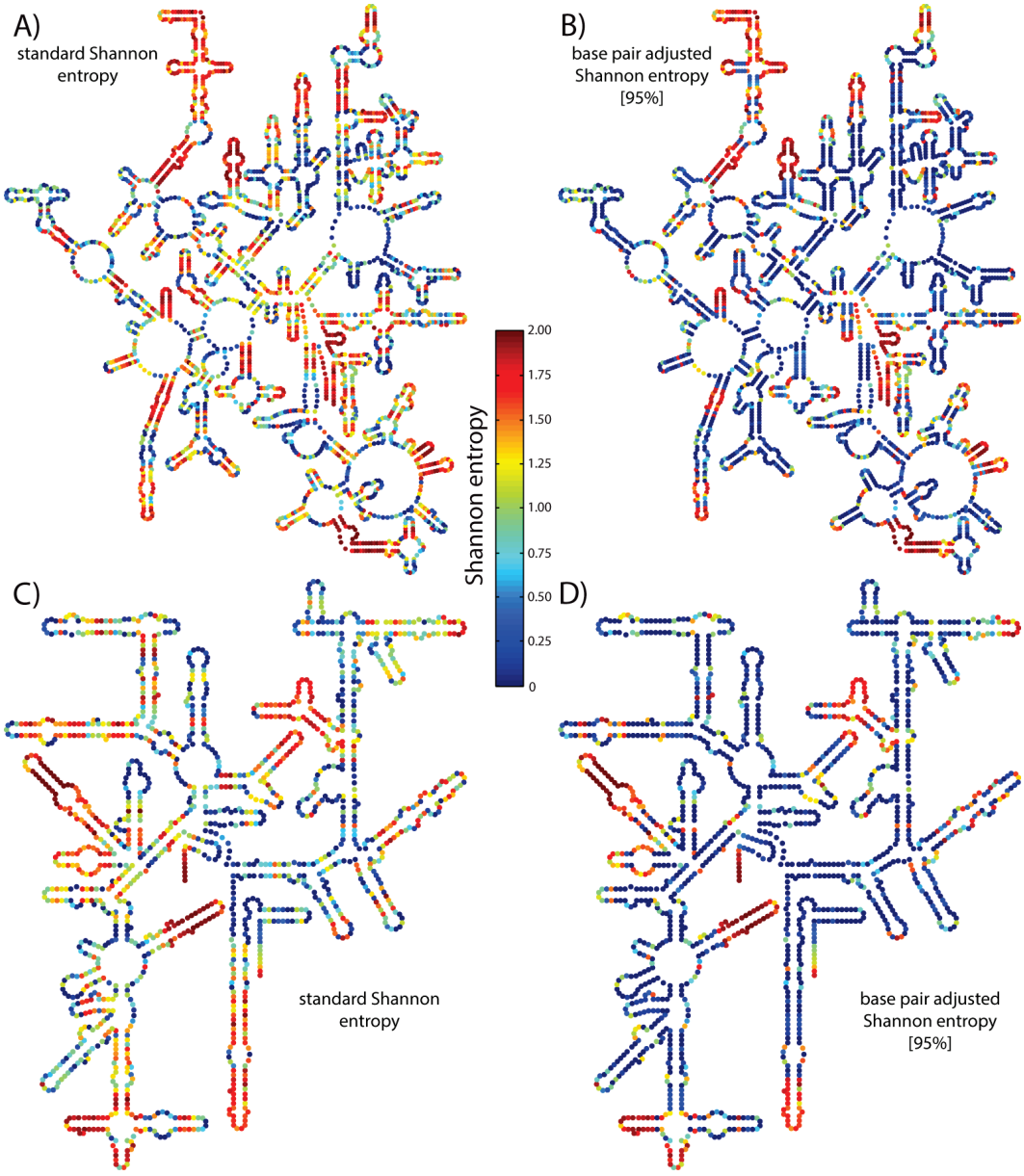

Figure 4.7. RiboZones alingment entropies mapped onto *E. coli* LSU and SSU secondary structures. Base pair adjusted entropies have been artificailly doubled to put them on the same scale as individual entropies. The 5S rRNA has been ommitted. A) LSU rRNA with all individual entropies. B) LSU rRNA with base pair adjusted entropies. C) SSU rRNA with all individual entropies. D) SSU rRNA with base pair adjusted entropies.

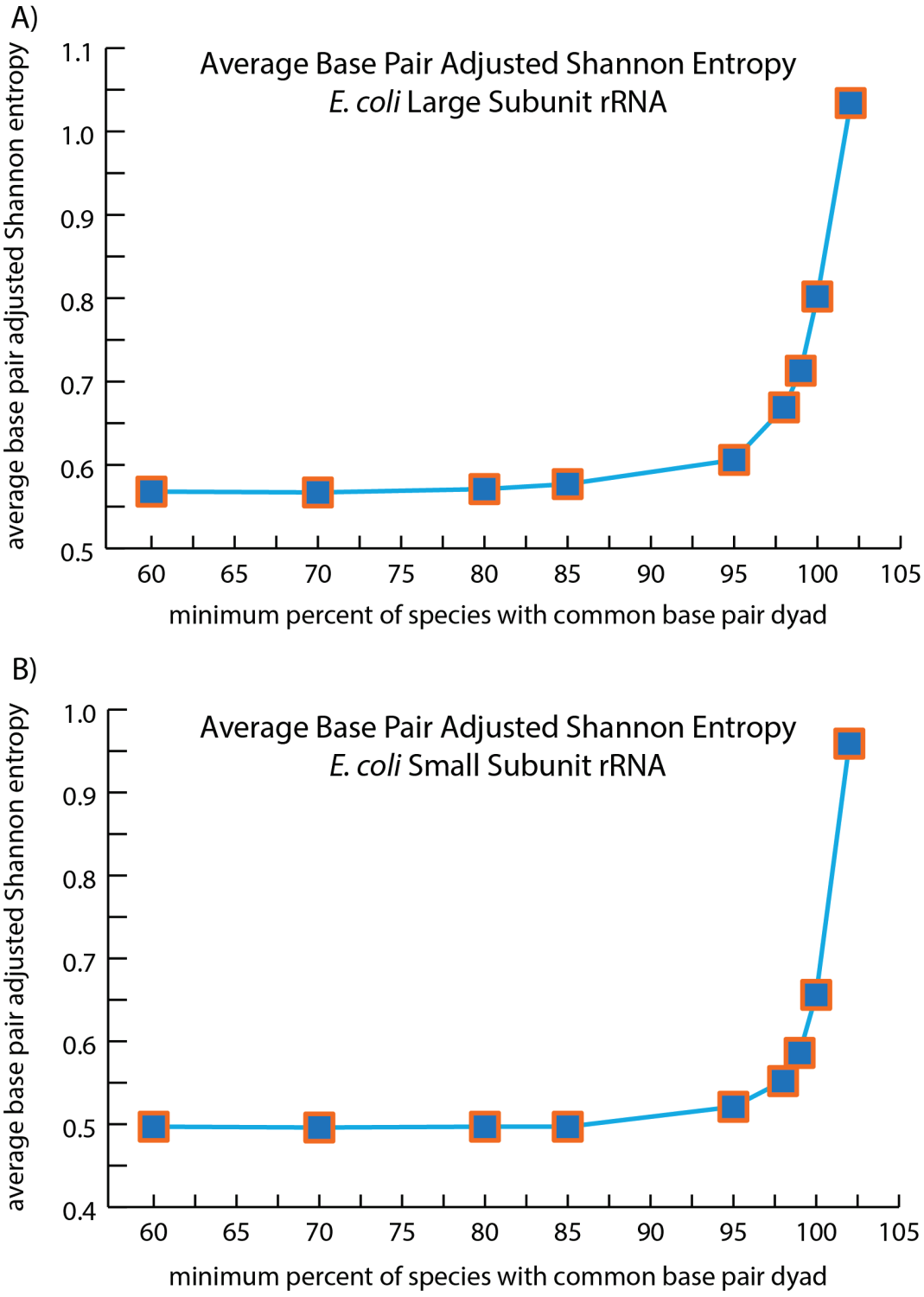

Figure 4.8. Average base pair adjusted Shannon entropy as a function of base pair cutoff percentage. Individual entropies were replaced with base pair entropies for positions where an allowed base pair dyad occurred the minimum percentage of the time. A percentage above 100% is equivalent to using individual entropies only. A) For *E. coli* LSU. B) For *E. coli* SSU.

4.5 Alignment Problems

Over the course of manual alignment correction, several major problems were noticed. The biggest problem is SILVA specific, the failure to align Domain I in many eukaryotes. To better illustrate this problem, screenshots of a subset of the alignment were taken. The alignment for the end tip of LSU Helix 10 for a subset of 11 species is shown (**Figure 4.9**). The SILVA alignment is shown in **Figure 4.9A**. For the prokaryotes, the first strand of Helix 10 is visible. However, two out of three eukaryotes, *Caenorhabditis elegans* and *Homo sapiens* do not have Helix 10 rRNA in this position. In Eukaryotes, Helix 10 is where the 5.8S and the 26S bind, so Helix 10 does not have a RNA loop. SILVA incorrectly places a large portion of Domain I for certain eukaryotes in the region where the prokaryotic Helix 10 tetra-loop would go. In Eukaryotes, this region is where Internally Transcribed Spacer 2 (ITS2)[82] would be. SILVA has ITS2 sequences in their database and incorrectly recognizes some rRNA as ITS2. This can be seen in **Figure 4.9A**, for *H. sapiens*. The lowercase sequence is not properly recognized and should be aligned with Helix 1. The sequence for *Caenorhabditis elegans* begins further along the alignment.

In 3D, the alignment problems are clearer. **Figure 4.10A** contains a 3D representation (red) of the portion of *E. coli* Helix visible in **Figure 4.9**. **Figure 4.10A** also contains a 3D representation (green) of the portion of *H. sapiens* rRNA that the SILVA alignment implies is homologous with the *E. coli* sequence. *H. sapiens* Helix 10 is included, along with hundreds of other nucleotides, which are incorrectly placed. The problematic eukaryotic sequences begin aligning correctly further downstream, but not

until Helix 25a starts at the beginning of Domain II. **Figure 4.10B** contains the same thing but on the 2D secondary structure.

Figure 4.9. Partial multiple sequence alignment for Helix 10 of the LSU. Only 11 species are shown for visualization purposes. A) SILVA alignment, B) RiboZones alignment, C) theoretical perfect structure based alignment.

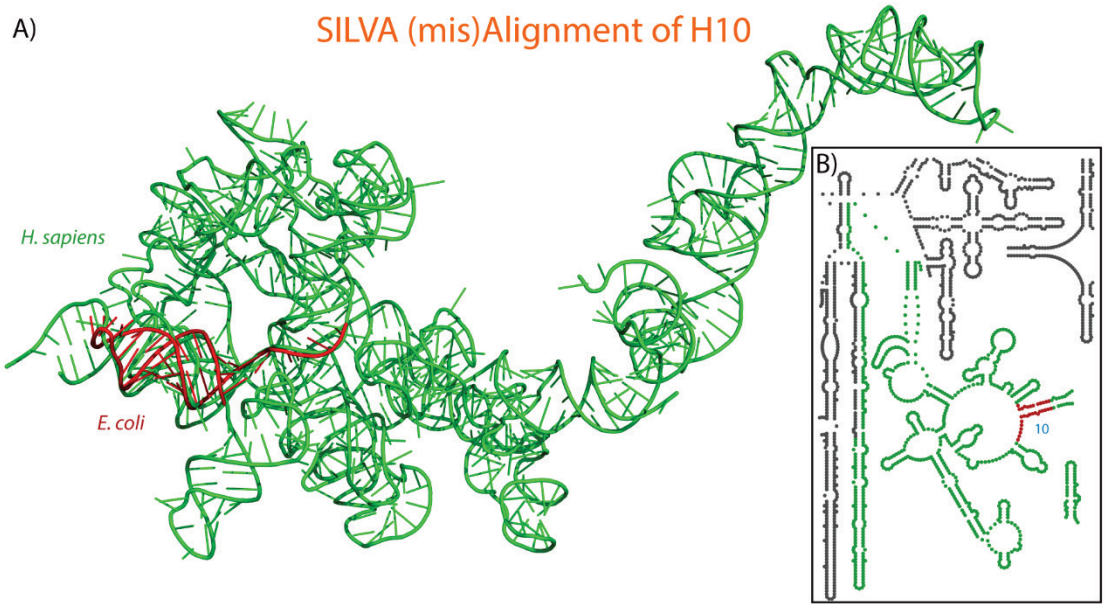

Figure 4.10. Molecular 3D representations of Helix 10 as shown in Figure 4.9. A) *E. coli* Helix 10 is shown in red. The predicted Helix 10 of *H. sapiens* , as predicted by the SILVA alignment, is shown in green. There is much more rRNA included than should be, because SILVA put most of Domain I inside the Helix 10 region for some of the eukaryotes. B) Partial secondary structure of *H. sapiens* rRNA. The misalignned H. sapiens rRNA is shown in green. The rRNA of *H. sapiens* that should be aligned with *E. coli* is shown in red.

The RiboZones alignment is shown in **Figure 4.9B**. Here, the first strand of Helix 10, the tetra-loop for prokaryotes, and the second strand of Helix 10 are visible. The RiboZones alignment has solved this problem by using MAFFT to add the problematic eukaryotic sequences to the SILVA alignment instead of using SINA to align those sequences. **Figure 4.9C** shows proposed further improvements to the alignment, which were not made to the RiboZones alignment.

The next major misalignment problem, of Helix 30, is shared by SILVA and CRW and is indicative of a major problem in the seed alignments and the expert understanding. Not all species have a Helix 30, so for those that don't, the Helix 31 rRNA gets misaligned into the Helix 30 rRNA causing inaccuracies for both Helix 30 and

Helix 31. SILVA's alignment is in **Figure 4.11A**. The alignment, for bacteria, starts with 5 nucleotides, presumably base baired, then continues with a 3 member RNA loop. The archaea have a 4 nucletide stem, and a 4 or 3 member RNA loop. The eukaroyotes have no dicernable pattern. Examing the secondary and tertiary structures available reveals this is a false representation of the data. A 3D representation of the alignment in **Figure 4.11A** for *E. coli* and *S. cerevisiae* is in **Figure 4.12A**. *E. coli* is in red, with the *S. cerevisiae* RNA in blue. **Figure 4.12C** has the same data projected onto the secondary structure of *S. cerevisiae*. It is evident that these are not the same helix.

The Ribozones alingment is shown in **Figure 4.11B** and **Figure 4.12B**. This alignment was created by manually moving appropiate columns to where it appears they should be. Bacteria have 5 base pairs, visible on both sides of Figure 12B. They also have a 5 member pentaloop in the center. Archaea also have 5 base pairs, but with a 4 member tetraloop. Eukaryotic sequences have a longer helix as this is the site of the small ES9. Eukaryotic rRNA loops vary, with *S. cerevisiae* having a tetraloop and *H. sapiens* having a hexaloop.

Similar problems with misalinging rRNA with the wrong helix also occurs in the regions of H15 and H31a of the LSU, and h9 and h21 of the SSU. In addition to these, less severe but significant rRNA shifts occur in H38, H60, h39, h33, and h44.

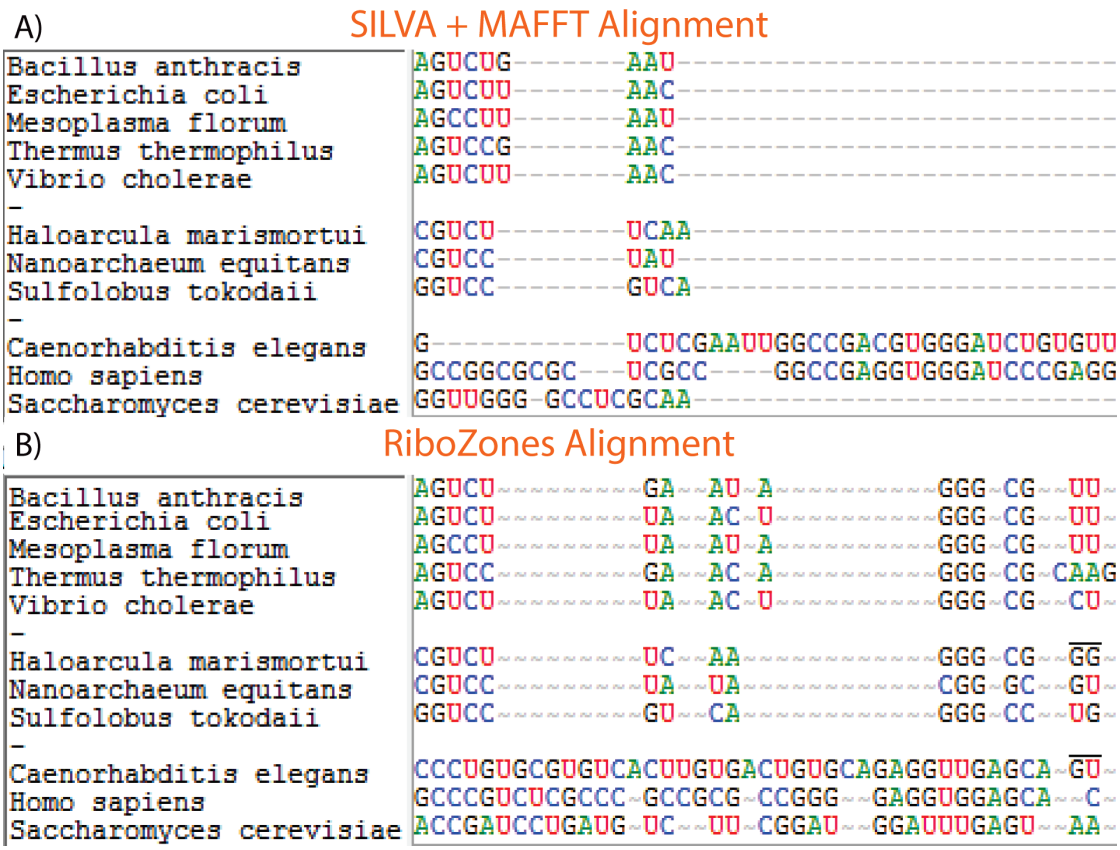

Figure 4.11. Partial multiple sequence alignment for Helix 31 of the LSU. Only 11 species are shown for visualization purposes. A) SILVA + MAFFT alignment, B) RiboZones alignment.

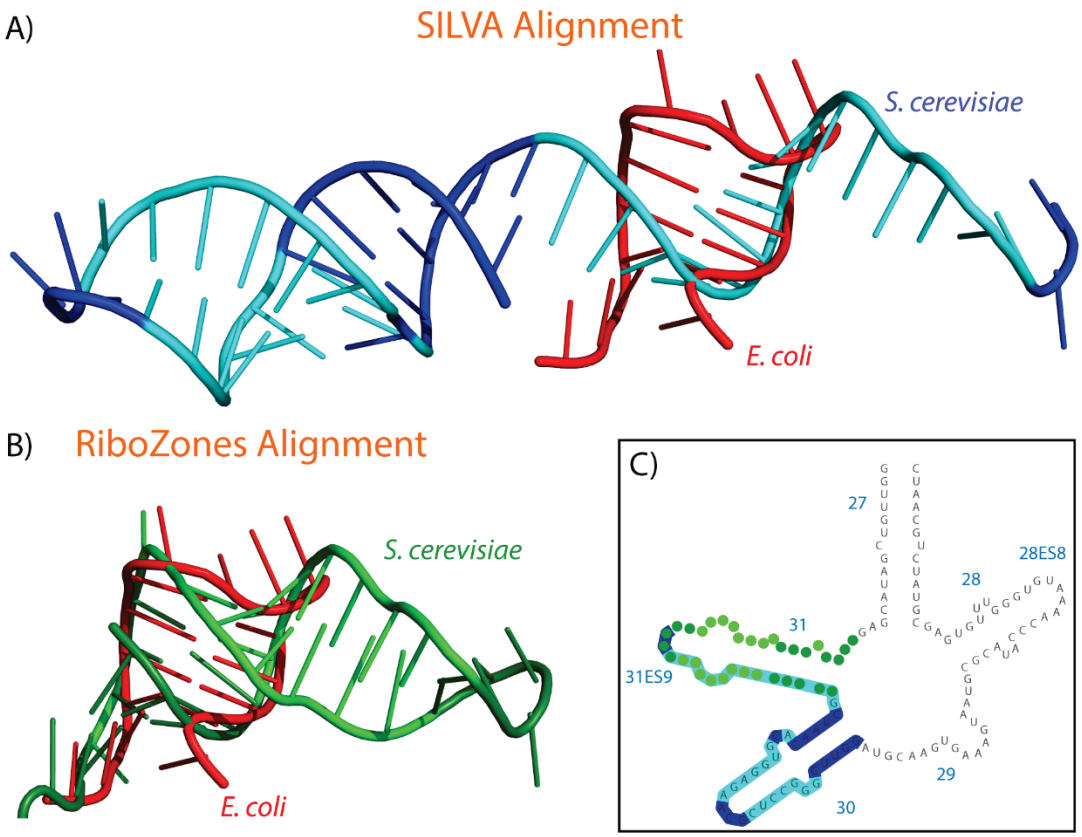

Figure 4.12. Molecular 3D representations of Helix 31 as shown in Figure 4.11. *E. coli* Helix 31 is shown in red. A) The predicted Helix 31 of *S. cerevisiae*, as predicted by the SILVA alignment, is shown in blue. The parts of *S. cerevisiae* H31 that should align with *E. coli* H31 are dark blue. There is much more rRNA included than should be, evidence of a poor alignment. B) The predicted Helix 31 of *S. cerevisiae*, as predicted by the RiboZones alignment, is shown in green. The parts of *S. cerevisiae* H31 that should align with *E. coli* H31 are dark green. The RiboZones alignment is correct here. C) The same data as in A and B, but on the secondary structure of *S. cerevisiae* instead of in 3D. The blue contour line highlights the same RNA as in A. The green circles highlight the same RNA as in B. Here, it is clear that H30 and H31 are distinct helices and should not be in the same alignment positions.

4.6 Alignment Comparison

Developing alignment algorithms is popular and is an ongoing effort.[50,83-85] How are the alignments compared and judged? For generic RNA aligners, the standard test dataset is BRaliBase II.[77] There are many different scoring algorithms and researchers still develop new scoring algorithms.[50,63,86] However, ribosomal RNA tends to be an exception and does not behave like normal RNA.

There is no formal test dataset for LSU nor SSU rRNA. The SILVA seed alignment and the CRW alignments are the closest things that the rRNA community has to a standard alignment. However, as shown, these standard alignments have some serious inaccuracies. The RiboZones alignment is an improvement with respect to the statistics studied here.

When judging alignments, the first variable noticed is the length of the alignment. **Table 4.1** and **Table 4.2** contain the length and other basic statistics for the LSU and the SSU respectively. The RiboZones alignment is compared to CRW, SILVA, MAFFT, and Clustal Omega. A quality rRNA alignment should be intermediate in length. Shorter alignments, like Clustal and MAFFT over align the RNA, not sufficiently separating expansion segments and optional helices from the common core of the ribosome. Over alignment also produces local disturbances in the alignment. However, an overly long alignment is less than ideal. Long alignments are indicative of potentially homologous RNA not being recognized as such. In some cases, it may be desirable for highly divergent RNA to be separated in the alignment. However, for the purposes here, as much homology as reasonable should be forced.

Table 4.1. Alignment statistics for the LSU for several alignment algorithms. The statistics are loosely correlated with completeness of the sequences and overall alignment quality.

	Clustal Omega	SILVA	MAFFT	SILVA + MAFFT	RiboZones	CRW*
Total Length [nt]	7108	10562	7460	9525	9435	14240
Gap Density [%]	55	70	57	67	66	78
Positions (no gaps)	1792	1857	1923	1950	2011	1157
Positions (few gaps)	2177	2368	2450	2453	2465	2276
Positions (highly conserved)	695	712	719	749	746	801
Positions (universally conserved)	264	242	272	327	282	102

Table 4.2. Alignment statistics for the SSU for several alignment algorithms. The statistics are loosely correlated with completeness of the sequences and overall alignment quality.

	Clustal Omega	**SILVA**	**MAFFT**	**RiboZones**	**CRW***
Total Length [nt]	2708	3444	2676	3152	8716
Gap Density [%]	41	54	41	50	82
Positions (no gaps)	1171	1151	1217	1260	0
Positions (few gaps)	1312	1323	1355	1364	961
Positions (highly conserved)	444	438	440	443	394
Positions (universally conserved)	222	218	216	217	0

A good example of an under alignment is Helix 98 of the LSU. RiboZones attempts to align each side of *E. coli* Helix 98 with the corresponding bases in *S. cerevisiae* Helix 98 (**Figure 4.13A**). In 3D (**Figure 4.14A**), these helices initially do not look homologous, but base pairs can be matched up manually. **Figure 4.14B** shows which parts of *S. cerevisiae* H98 are approximately homologous with *E. coli* H98. The CRW alignment has no overlap between the prokaryotic H98s and the eukaryotic H98s. Since some bacteria and some archaea do not have a H98, it is unknown whether LUCA had a H98. It is possible H98 evolved separately through convergent evolution. If these helices evolved separately, it would make sense for a phylogenetic based alignment to have these helices separated. However, the goal of RiboZones is to produce a structure-based alignment useful for comparing structure. Therefore, it is preferable to put these helices together.

Gap density is highly correlated with overall alignment length. Similarly, it should be an intermediate value. It is not very useful when comparing widely differing alignments.

The RiboZones alignment is best with respect to another useful metric, the number of positions with no gaps or very few (<5%) gaps, which is especially useful when in the sequence collection stage of building an alignment. The number of (nearly) complete positions is very sensitive to incomplete sequences. CRW's alignment, with about 10% incomplete sequences, produces a non-gapped position count of 1157 for the LSU, and surprisingly, 0 for the SSU (**Table 4.1** and **Table 4.2**). The low gap counts are 2276 for the LSU and 961 for the SSU. These numbers rise when using the RiboZones

sequence list though any of the alternative alignment methods. The RiboZones alignment has the highest numbers.

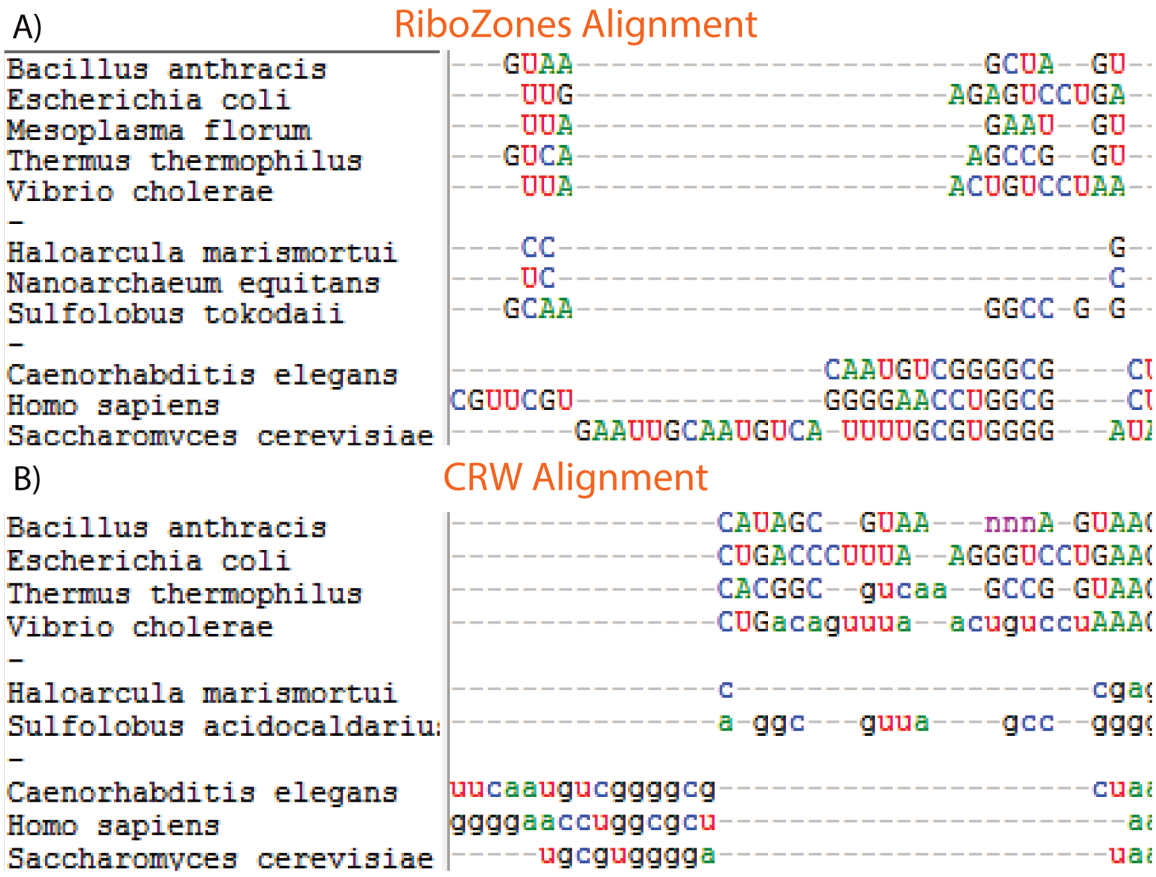

Figure 4.13. Partial multiple sequence alignment for Helix 98 of the LSU. Only 9 species are shown for visualization purposes. A) RiboZones alignment, B) CRW alignment.

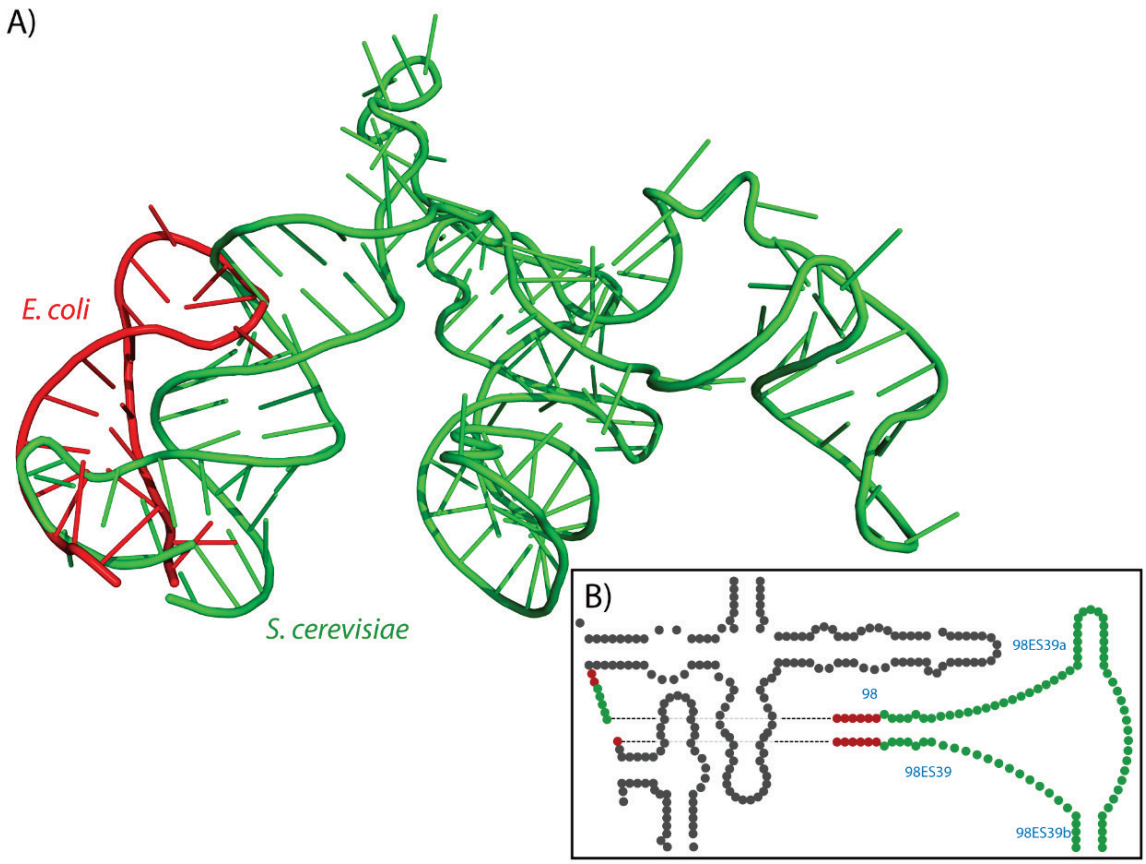

Figure 4.14.. Molecular 3D representations of Helix 98 as shown in Figure 4.14. A) *E. coli* Helix 98 is shown in red. Helix 98 of *S. cerevisiae* is shown in green. The *E. coli* H98 is partially homologous with the *S. cerevisiae* H98 if rotated and translated. The CRW alignment shows no homology. This is evidence of under alignment. B) The same data as in A, but on the *S. cerevisiae* secondary structure. H98 and its expansion segments are shown in green. The part of *S. cerevisiae* H98 homologous with *E. coli* H98 is shown in red.

Similarly, the number of positions, which are highly or universally conserved, is a useful indicator of both sequence completeness and alignment quality. Incomplete sequences and misalignments prevent these numbers from reaching their unknown true value. Care must be taken when interpreting these numbers. The correct structure-based alignment would have a somewhat lower score than a purely sequence based method. Sometimes, preserved sequence has physically moved in structure between the different domains of life. For example, a base pair could be inserted at the beginning of a helix, which would shift the remainder of that helix in the alignment. Another example would be a nucleotide, which was at the beginning of a loop in one species, but evolved a binding partner and becomes the end of the helix instead. The SILVA and CRW alignments reflected the majority of these events, showing an improvement over Clustal, but some were overlooked. The RiboZones alignment has been adjusted to reflect these changes in more locations.

Additional statistics become possible when filtering an alignment down to a specific species. Without doing so, calculating average entropy would be dominated by the gaps. The alignments are between 41% and 82% gaps (**Table 4.1** and **Table 4.2**). The variation of Shannon entropy used here is using a prorate strategy for counting the gaps. Each gap is treated as if it was an A, C, U, or G simultaneously with 25% probability each. With such a high gap penalty, the overall average Shannon entropy for these alignments is approximately 1.8. Additionally, without filtering to a specific species, there is no way to use specific known base pairs in analysis. Covariation analysis could be done to predict base pairing, but RiboZones uses measured known base pairs from 3D structures.

Here, the alignment is filtered onto *E. coli*, but the methodology could be repeated onto any species for which a secondary and tertiary structure is known. Only positions that are present in *E. coli* are included in the stats in **Table 4.3** and **Table 4.4**.

Average entropy is the average gap pro-rated regular Shannon entropy over the *E. coli* filtered alignment. All the alignments tested produced similar average entropies. The CRW alignment scored slightly lower than the others score, but does not use the same species list, so a direct comparison is not fair. The CRW contains many repeats of similar sequences and lacks some of the diversity in the RiboZones alignment.

Average entropy (adjusted) is the average base pair adjusted Shannon entropy over the *E. coli* filtered alignment. The pure SILVA + MAFFT for the LSU scored the lowest (best), with the RiboZones manually adjusted alignment being extremely close. Most of this 0.01 difference is likely to be caused by attempts to manually adjust H15, H16, H17, and H18. These regions are difficult and will need care to improve in a future version of the alignment. For the SSU, the RiboZones alignment received the best score. In both cases, the RiboZones alignment outperformed the CRW alignment, despite the CRW having a lower entropy to start with. This suggests that the RiboZones alignment contains a higher number of correctly aligned base pairs, partly due to the actual alignment but mostly due to the higher completeness of the dataset.

Table 4.3. Alignment statistics for the LSU for several alignment algorithms. For these statistics, the alignment was first filtered down to only the positions in the alignment corresponding to positions in *E. coli*. The statistics are strongly correlated with completeness of the sequences and overall alignment quality.

	Clustal Omega	**SILVA**	**MAFFT**	**SILVA + MAFFT**	**RiboZones**	**CRW***
Average entropy	1.08	1.06	1.05	1.03	1.03	0.96
Average entropy (adjusted)	0.81	0.66	0.73	0.60	0.61	0.69
Structural divergence (rmsd) [Å]	20.82	9.45	19.81	9.45	7.84	7.79
Structural divergence (rmsd, common) [Å]	19.65	6.80	13.50	6.80	6.44	6.75

Table 4.4. Alignment statistics for the SSU for several alignment algorithms. For these statistics, the alignment was first filtered down to only the positions in the alignment corresponding to positions in *E. coli*. The statistics are strongly correlated with completeness of the sequences and overall alignment quality.

	Clustal Omega	**SILVA**	**MAFFT**	**RiboZones**	**CRW***
Average entropy	0.95	0.97	0.96	0.96	1.07
Average entropy (adjusted)	0.61	0.53	0.60	0.52	0.92
Structural divergence (rmsd) [Å]	18.88	13.24	13.77	11.18	8.32
Structural divergence (rmsd, common) [Å]	9.17	8.37	7.93	8.09	8.22

Finally, a fully structure based statistic, structural divergence, is introduced. Structural divergence has been used as an integral tool in comparing and adjusting alignments. For an overall comparison, the root-mean-square-deviation (RMSD) of the structural divergence is calculated and shown in **Table 4.3** and **Table 4.4**. Clustal and MAFFT had the least performance. The RiboZones alignment improved over the SILVA based alignments significantly, achieved mostly by fixing the incorrect helices previously mentioned.

The CRW alignments appear to have outperformed RiboZones' alignments however, this is misleading. CRW has been shown to under align some helices, particularly those that are most structurally diverse. Since these helices are not aligned in CRW, there is no corresponding RMSD for them, biasing the CRW statistic to be lower. To remove such bias, the RMSD of the common areas was calculated. To calculate this, only the positions of *E. coli*, which all alignments say something in *S. cerevisiae* aligns with it, were included in the RMSD calculation. With the bias removed, RiboZones gets the best RMSD score.

4.7 Discussion

There are several major problems with SILVA/SINA. There are sequence level issues, such as incomplete sequences and improper detection of junk RNA. This is most evident when studying LSU eukaryotes. Some sequences correctly contain 5.8S rRNA and no ITS2. Many sequences do not contain 5.8S rRNA at all. Many sequences incorrectly contain ITS2. Many LSU sequences are missing Domain I, Domain VI or both. In addition to ITS2 inclusion, there are other spacers such as the IVS region of

Helix 9 in certain species incorrectly included. The sequence level problems cause alignment level problems.

SILVA has the most issues in Domain I, particularly with aligning 5.8S sequences. The seed alignment appears to have inaccuracies propagating to the produced alignment. SILVA does not recognize some 5.8S sequences as valid 5.8S sequences, and puts them in an unaligned region of the alignment. Sometimes it aligns other 5.8S sequences to where ITS2 should go. Domain I contains the genuinely variable and difficult to align Helices 15, 16, 17, and 18. Helix 15 is not properly separated from Helix 16 causing both to misalign. In certain species with 5.8S problems, there is misalignment up through Helix 25, causing the entire Domain I to come out incorrectly for those species.

SILVA is more accurate with the other Domains. Domain II's problems are less severe than Domain I. The major problem is that Helix 30 is not properly separated from Helix 31. There are also misalignment issues in Helix 38. CRW's alignment shares many of the same problems. Domain III is genuinely difficult to align because it is structurally variable. Making a master alignment of Domain III would be challenging. Improving on the alignment for a subset of species would be possible, but only through more advanced alignment techniques involving known secondary and possibly tertiary structures of most species. Domain IV and V are highly conserved and as such align quite well. Minor adjustments need to be made, but the major features are all correct. Domain VI has moderate problems. A larger than expected number of manual corrections needs to be made for a relatively conserved domain. Helix 98 does not align well, caused by the huge expansion segment in this region.

The Small Subunit rRNA alignment has similar problems. Helices 9 and 21 are improperly aligned. The main helix is not properly separated from the expansion segments. This also happens in h33 to a minor extent. Helices 39 and the end of h44 also needed manual adjustment.

A database like SILVA is an excellent idea. "**SILVA** provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (*Bacteria***,** *Archaea,* **and** *Eukarya*)."[87] While this statement is true, there is room for improvement. SILVA admits their accuracy problems on their documentation page. "We would rate our SSU SEED alignment for all *Bacteria* and *Archaea* as good and for *Eukarya* as reasonable." "The LSU alignment […. for] *Bacteria* and *Archaea* could be rated as good. The *Eukaryotes* need definitely further attention." [88] Hopefully, the further attention will be executed soon. The issues documented here cause SILVA alignments to need more manual curation then they claim. It would be beneficial to the community to have these issues documented on the SILVA site, and for collaborations to begin to fix up the remaining issues. It would be effort well placed.

The SILVA alignments are a good starting place for making an rRNA alignment. Ultimately, they achieved quality on par with or exceeding the CRW alignments. The major difference is that SILVA is actively being maintained and improved and CRW is not. SILVA would gain more popularity when the issues are fixed.

In the meantime, the RiboZones alignment is at least as good of an alignment in most places, with better performance in some other places. Someone needing to make his or her own rRNA alignment would be well served to start with the RiboZones alignment.

The alignment could be used as a template alignment with programs such as MAFFT to add new and different sequences. The alignment could then be manipulated to meet their individual research needs. The RiboZones alignment should be the standard rRNA template alignment until the SILVA alignment can be fixed.

Two statistics are shown to be useful, base pair adjusted entropy and structural divergence. RiboZones software was used to calculate and visualize these statistics, but there is no current feature of RiboZones to aid in the actual editing of the alignment. Future improvements to RiboVision may include a built in alignment editor and more features helpful to simultaneously studying multiple structures.

The development of an alignment algorithm that automatically uses base pair adjusted entropy and structural divergence would be quite beneficial. Such an algorithm could also incorporate RNA folding algorithms to help predict the secondary and tertiary structures of rRNA helices. Such algorithms would need adjustments to be able to handle the more special case of rRNA versus regular RNA. Until the availability of a structure aware rRNA alignment program specialized for the special needs or ribosomes, template based methods must be used.

While the RiboZones alignment is not perfect, it is high quality relative to the alternatives of SILVA and CRW. Visualizing the alignment using RiboVision and base pairing entropy, allows the study of the ribosome on a helix-by-helix basis. Questions such as which parts of the ribosome are universally in common can begin to be answered. Comparative sequence and structure analysis can lead to a deeper understand of not only the structure of the ribosome, but its evolution and function. The focus of the next chapter is using this alignment to study the structure of the ribosome more deeply.

CHAPTER 5

UNDERSTANDING RIBOSOMAL STRUCTURE DIVERGENCE FROM THE COMMON CORE HAS EVOLUTIONARY IMPLICATIONS

5.1 Introduction

Understanding the origin of translation is central to understanding the origin of life itself. Most of the translation system is universal across all life on Earth. The translation system is a multiple component machine, composed of complex interactions between both RNA (mRNA, tRNA, and rRNA, etc.) and proteins (rProteins, initiation factors, elongation factors, aminoacyl tRNA synthetases [aaRS], etc.). The ribosome is the central component of the translation system. The translation system was bootstrapped out of a prebiotic world. Our goal is to build models of ribosomal origins and evolution.

Comparative structure analysis, combined with comparative sequence analysis, and evolutionary principles, allows the rational construction of a parsimonious model of ribosomal evolution. The ribosome had most of its current functionality at the time of the Last Universal Common Ancestor (LUCA). We need to reverse engineer LUCA to understand its origins. Fortunately, the structure of the ribosome is a molecular fossil, containing evidence of its own evolution. A mixture of top-down and bottom-up approaches is ideal.

We seek to define the universal common core, which is composed of the elements of the ribosome that are common to all living systems, within some well-defined statistical limit. We also define domain-specific common cores, which are common to each of the three domains of life. This information will be useful in

- defining and characterizing LUCA,

- rooting the tree of life,
- determining functionally important elements of the ribosome,
- determining the functions of domain-specific elements,
- establishing pathways and mechanisms of ribosomal evolution,
- establishing the origin of the domains of life,
- fully characterizing the differences between the domains.

The availability of atomic-resolution structures of ribosomes from species across the entire biodiversity of life provides data for comparative analysis and validation of models against additional species. Here, we establish the common cores and the pathways of ribosomal evolution since LUCA. Later in Chapter 6**,** this pattern is used to model pre-LUCA ribosomal evolution, rewinding the tape of life.

5.1.1 Common Core

Early sequence analysis predicted a common core of both LSU and SSU rRNA.[89,90] Sequence analysis of many genes reveals blocks of sequence that are either (i) universally present in all life, (ii) highly conserved with one or two domains of life, or (iii) variable amongst all life. rRNA genes are not an exception to this rule. In fact, because the SSU rRNA gene follows this rule so well, without significant horizontal gene transfer, it is the standard gene used for classifying new species and building phylogenetic trees. *E. coli* LSU and SSU secondary structures were the first to be determined in 1980 and 1981.[91,92] For this obvious reason, every other species was compared to *E. coli*. Sequencing of the mouse 28S rRNA gene, revealed that most of the rRNA had the same secondary structure as *E. coli*, with the extra nucleotides in

eukaryotic sequences being localized to certain regions.[90] Further sequence and experimental information allowed secondary structure models to be refined further.[89,93-97]

The availability of multiple high-resolution 3D structures of ribosomes confirmed the presence of a common structural core, and established that the previously predicted secondary structures were approximately 97% accurate[56,58,98-100] Eukaryotic ribosomes have evolved by growing rRNA expansion segments at the subunit surfaces.[101] The role of eukaryotic expansion segments is not well understood, but they likely play a role in regulation and initiation.[99,102]

5.1.2 Bacterial / Archaeal Divergence

Ever since Woese's landmark paper on the three domains of life,[103] there has been great debate on the origins of the three domains.[104,105] The current consensus is that bacteria and archaea split off from LUCA around 3.5 billion years ago.[106-108] Analysis has been done comparing the bacterial and archaeal ribosomes.[6,58,109,110] The differences in the rRNA are relatively minor, but archaea and bacteria share only about half their proteins. It is currently unknown why Archaea and Bacteria split. It is also unknown how the differences in the rRNA are related to the differences in other aspects of archaea such as the translation initiation system and the different membranes. Precisely determining the common core of bacteria and archaea will aid in elucidating their differences. Conserved structural differences can be correlated with each other, with rProtein changes, and with changes in accessory machinery.

5.1.3 Detailed Common Core

The RiboZones structure-based alignment is used to define the common core for both the LSU and the SSU rRNAs. We have defined the common core as any part of the

ribosome that is present in 95% of the species in the dataset. We have collected 133 fully sequenced organisms to represent the entire tree of life (See details in **CHAPTER 4.2**). At the 2nd level of refinement, we distinguish conserved base pairs, conserved single stranded residues, and non-conserved single stranded residues are made. Expansion sites are clearly visible. Bacterial and archaeal ribosomes will be compared. Finally, we establish a relationship between eukaryotic expansion segments, time, and organismal complexity.

5.2 Methodology

5.2.1 Phylogenetic Tree

LSU lengths and genome lengths for our species list were mapped onto a phylogenetic tree. Alignment data is the same alignment as generated in Chapter 4. LSU rRNA lengths were measured from the sequence data. We added a constant 120 to each one to approximate the length of the 5S. Genome c-values were retrieved from NCBI. Since prokaryotic rRNA lengths vary relatively little compared to eukaryotes,[100] the average rRNA lengths for bacteria and archaea are used to represent the whole domain. A phylogenetic tree was calculated from sTOL.[111] The initial visualizations were generated with iTOL.[112]

5.2.2 Fine-Grained Onion

Fine-grained onion data for representative ribosomes was mapped onto both secondary and 3D structures. Fine-grained onion data were calculated using RiboLab software. The fine-grained onion metric is the distance, in 3D, from a particular point in the subunit to each nucleotide. For simplicity, the distance of a nucleotide was calculated

from the position of its P atom. For the LSU, the origin point is chosen to be the peptidyl transference center, here approximated as the 03' prime of residue 76 from the p-site tRNA molecule in PDB ID 2J01. For the SSU, the origin point is the decoding center, here approximated around the phosphate position of A1493 in PDB ID 2j01.

5.2.3 Basic Common Core

Binary common core definitions, is common core or is not, are projected onto the model species, *E. coli*, *H. marismortui*, and *S. cerevisiae*. Common core definitions are generated using RiboLab and visualized using RiboVision. The alignment is first projected onto the model species, *E. coli*, *H. marismortui*, or *S. cerevisiae*. In additional to the universal common core, a prokaryotes only, and common cores for each individual domain of life are calculated. Positions that have a gap frequency greater than 5% are marked in black. The rest, the common core, is marked in an appropriate color. Since RiboZones does not have an alignment for 5S rRNA at this time, and the 5S does not vary much, the whole 5S was marked as common core. RiboLab produces a RiboVision custom dataset file containing a "DataCol" with entropy. Such data was used to define the binary "ColorCol" for these figures.

5.2.4 Detailed Common Core

More detailed definitions of the common core were projected onto *E. coli* and *H. marismortui*. Detailed common core data is generated using RiboLab and visualized using RiboVision. The alignment statistics are projected onto the model species, *E. coli* or *H. marismortui*. The prokaryotic only alignment filter was applied. Positions that have a gap frequency greater than 5% are marked in red or orange, except the 5' and 3' ends, which were marked in gray. Conserved base pairs were marked in dark blue. Conserved

single nucleotides, with a Shannon entropy approximately corresponding to 95% conserved were marked in medium blue. Non-conserved single nucleotides were marked in light blue. Since RiboZones does not have an alignment for 5S rRNA at this time, and the 5S does not vary much, the whole 5S was marked in gray. RiboLab produces a RiboVision custom dataset file containing a "DataCol" with entropy. Such data was used to define the "ColorCol" for these figures.

5.2.5 Ribosome size Timeline

A timeline of approximate ribosomal size as a function of evolutionary time was computed. The ribosomal timeline was created using the same data in the phylogenetic tree made here (**CHAPTER 5.2.1**, **Figure 5.1**). The major clade divergences represented by our tree were used as time points. To date a time point, the estimated time since divergence for that clade split was taken from TimeTree.[113,114] Each time point represents a common ancestor. The rRNA size of that common ancestor is approximated as the minimum rRNA size in our dataset for species which diverged after that time point.

5.3 Localizing Ribosomal Growth

Eukaryotic rRNA sequences are larger and have more variable lengths than prokaryotic rRNA sequences. The RiboZones multiple sequence alignment (MSA) contains only full-length sequences, so rRNA length measurements are precise enough to analyze with low sample size. Prokaryotic sequence length varies relatively slightly. Bacteria LSU rRNA has an average length of 3087 (stdev 84), and bacterial SSU rRNA has an average length of 1522 (stdev 28). Archaea LSU rRNA has an average length of 3019 (stdev 70), and archaeal SSU rRNA has an average length of 1484 (stdev 15).[100] There is greater variability among the eukaryotes. *S. cerevisiae*, the standard RiboZones single celled eukaryote, has an LSU length of 3674 and a SSU length of 1800.

LSU rRNA length correlates with organismal complexity, but not with genome size. To visualize any potential rRNA length trends, the sizes of all eukaryotic sequences in the dataset were mapped onto a phylogenetic tree (**Figure 5.1**). All eukaryotic sequences are significantly larger than the prokaryotic sequences. Most unicellular eukaryotes are on the low end of the eukaryotic range. Trypanosomes (*Trypanosoma. brucei* and *Leishmania major*) are anomalously large for unicellular organisms. The development of multicellularity correlates with larger LSU size. The SSU grows at a much slower rate. There is a prominent size increase in birds and mammals. The human LSU is the largest in the dataset, with a length of 5347. Human's closest relative in the dataset, chimpanzee, is 5125, over 200 nucleotides smaller.

As previously noted, rRNA growth is mostly localized to a few distinct helices, which are on the outer surface of the ribosome. To visualize this, using RiboZones tools, the distance of each nucleotide from the center of their respective subunits was mapped

on the secondary structure using RiboVision (**Figure 5.2** and **Figure 5.3**). The sites of expansion are generally in the yellow, orange, and red zones, over 80Å from the center.

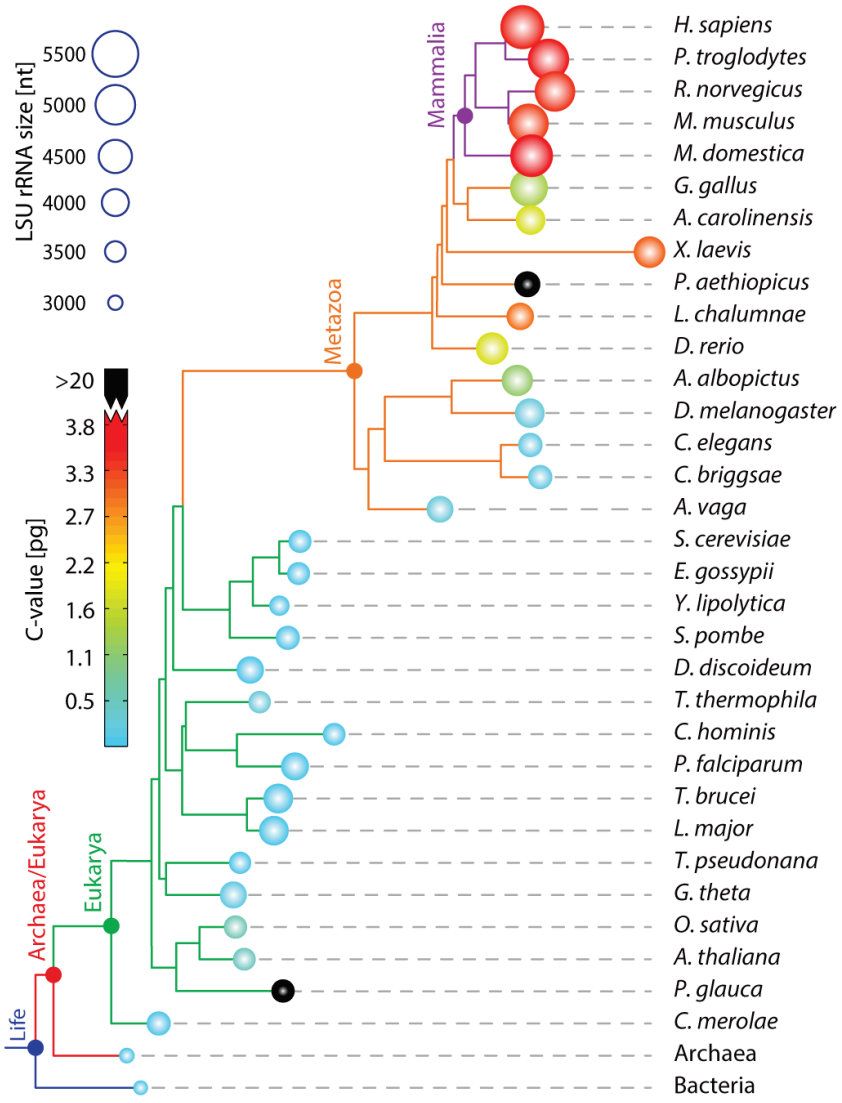

Figure 5.1. Phylogram indicating the sizes of LSU rRNAs and the sizes of genomes. Circle radii are proportional to total length of LSU rRNAs. Circles are colored by C-value, which is genome size measured in picograms. Two species here have anomalously high C-values and are colored in black (*P. aethiopicus*: C-value 133 pg, and *P. glauca*: C-value 24 pg). The sizes of archaeal and bacterial LSU rRNAs are highly restrained, so they are represented by just one species each. The phylogram was computed using sTOL[111] and visualized with ITOL [112]. Three species (*P. aethiopicus*, *A. vaga*, *P. glauca*) were manually added to the phylogram, because the genomes are not sufficiently annotated for sTOL analysis.

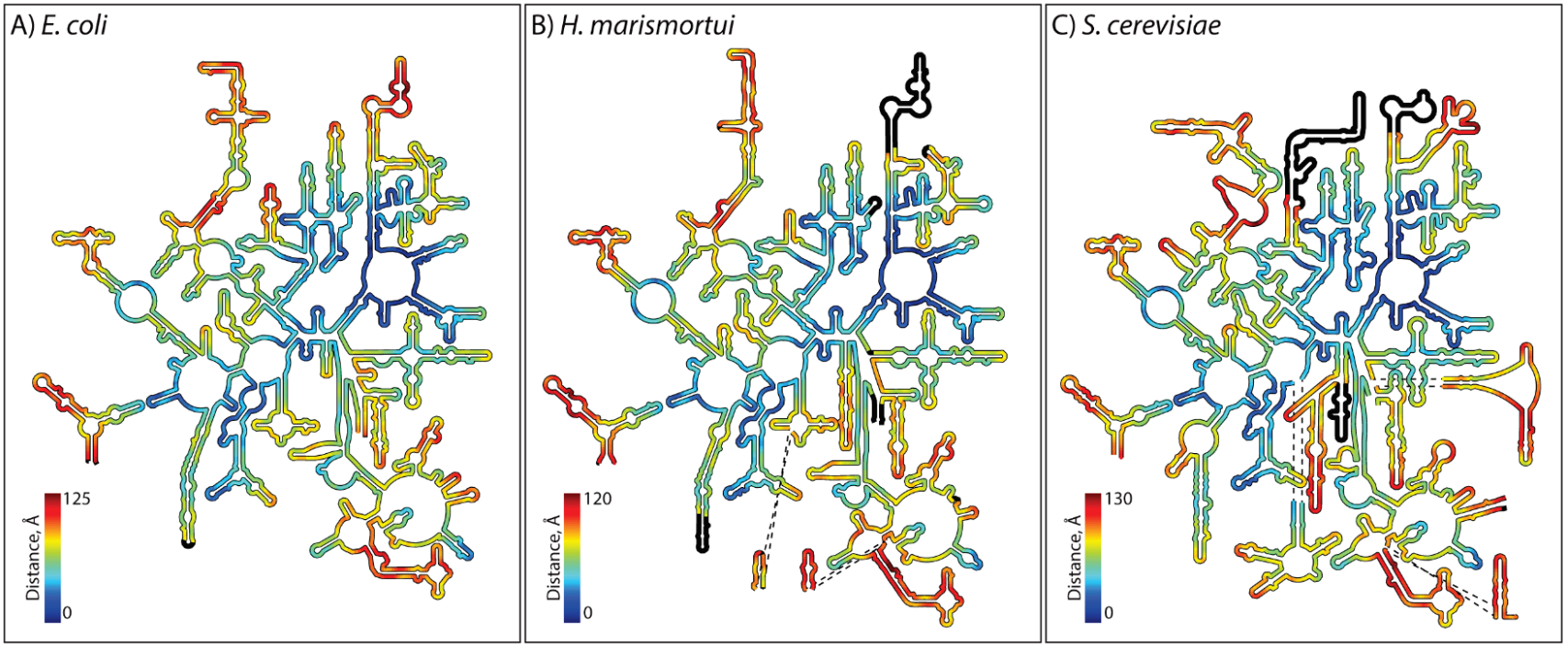

Figure 5.2. LSU rRNA secondary structures. A) *E. coli*, B) *H. marismortui,* and C) *S. cerevisiae.* The color indicates the proximity in three dimensions to the site of peptidyl transfer. Blue is close to the site of peptidyl transfer and red is remote. Nucleotides that were not experimentally resolved in three dimensions are black on the secondary structures.

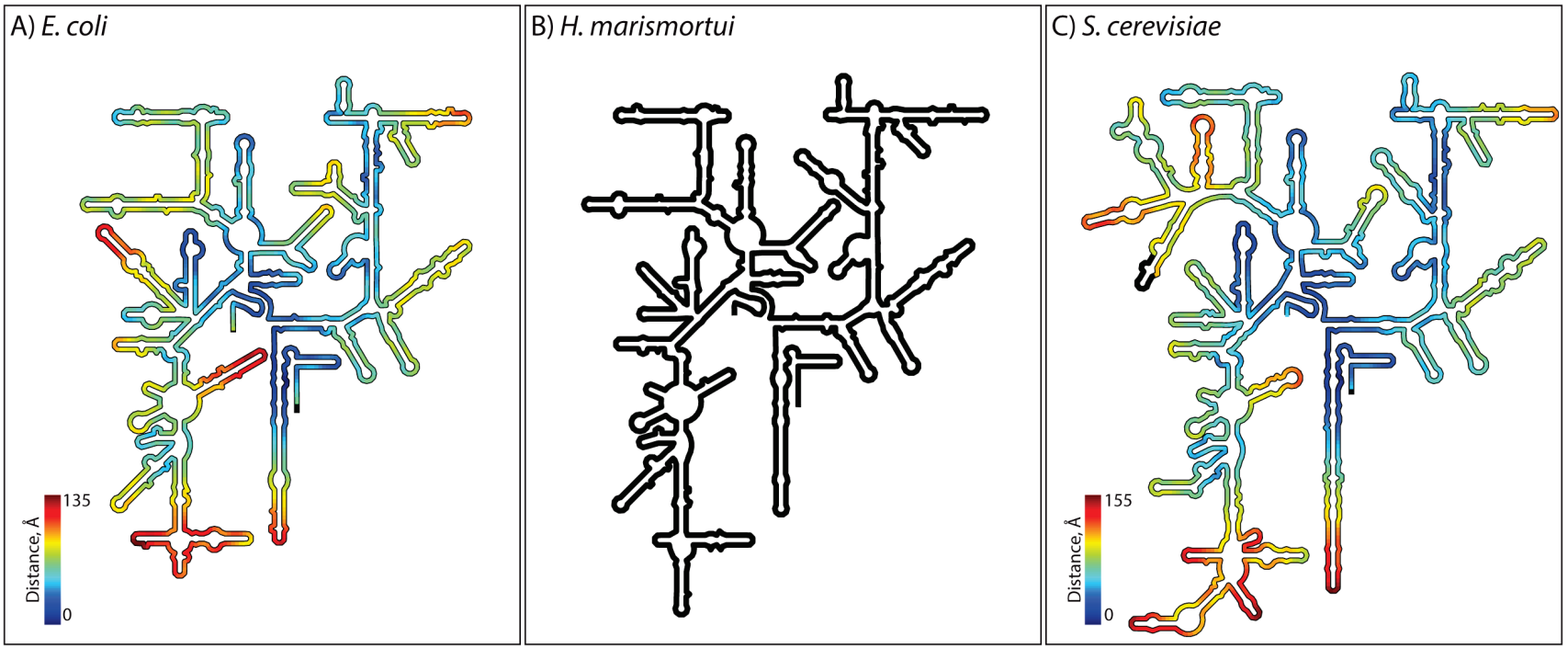

Figure 5.3. SSU rRNA secondary structures. A) E. coli, B) H. marismortui, and C) S. cerevisiae. The color indicates the proximity in three dimensions to the decoding center. Blue is close to the site of decoding and red is remote. Nucleotides that were not experimentally resolved in three dimensions are black on the secondary structures. There is no 3D SSU structure for H. marismortui or any other archaea available.

5.4 Defining the Common Core

Here, the common structure of ribosomes is defined on the individual nucleotide level. The first level definition of the "common core" is any nucleotide that is present in 95% of the sequences in the dataset. The broad definition allows variable sequence, but single-stranded nucleotides, to be included in the common core. A common core definition based purely on sequence conservation would exclude many of these single stranded regions, especially the RNA loops. RiboZones facilitates mapping an alignment onto a specific secondary structure. Species filters can be applied to make different versions of the common core. While calculating sequence entropy, positions exceeding a threshold gap frequency can be marked and visualized. Visual analysis is simplified though use of RiboZones' superimposed and templated secondary structures.

A comprehensive version of the common core (LSU: **Figure 5.4**, SSU: **Figure 5.5**) uses all three domains of life. We use the full RiboZones MSA, 133 species consisting of bacteria, archaea, and eukarya. Representative species for visualization are *E. coli* for bacteria, *H. marismortui* for archaea, and *S. cerevisiae* for eukaryotes. For each nucleotide in these species, Shannon entropy was calculated, using a threshold gap frequency of 5%. For simpler presentation, as an alternative to Shannon entropy, all nucleotides under 5% gap frequency are colored and all nucleotides above 5% gap frequency are black.

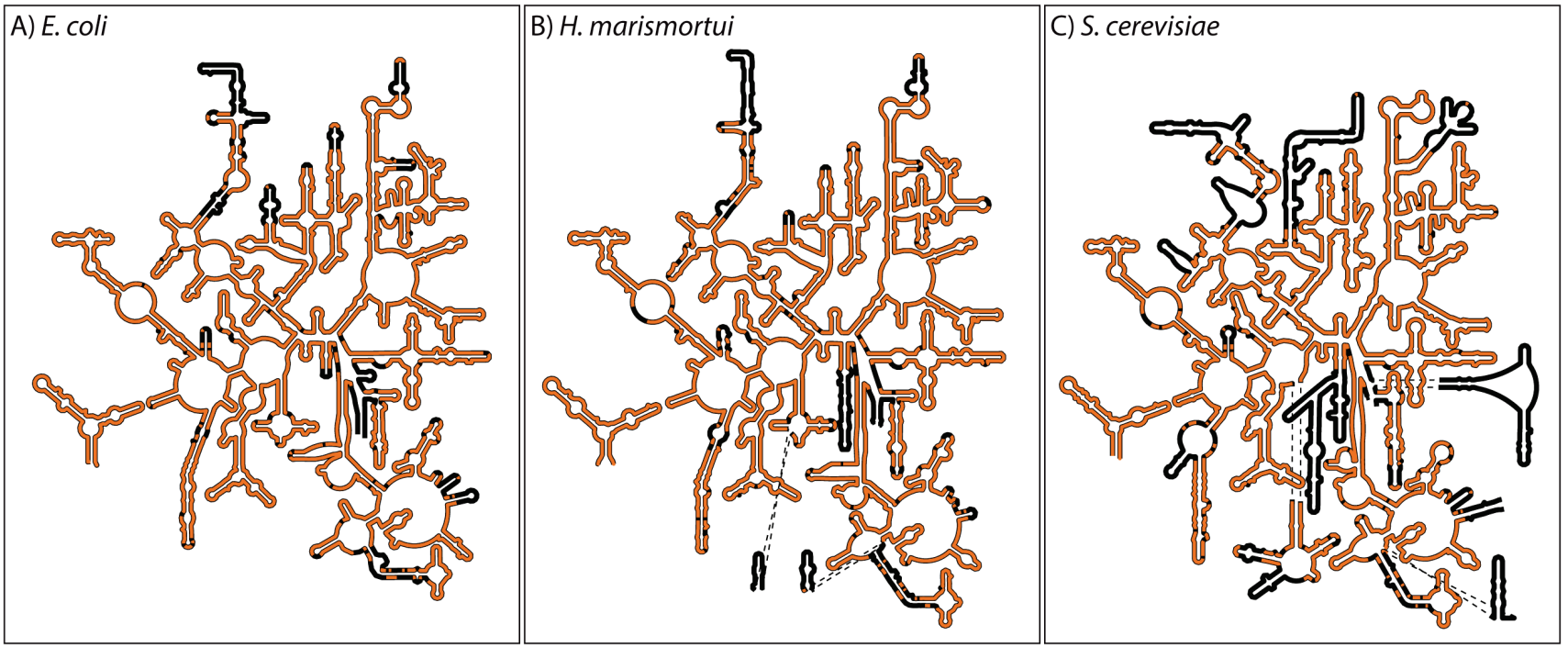

Figure 5.4. LSU rRNA secondary structures. A) *E. coli*, B) *H. marismortui,* and C) *S. cerevisiae.* The color indicates presence in the common core. Orange is included in the common core and black is excluded. Nucleotides are included if they are present in 95% of the species in the whole RiboZones alignment. Most black areas are helices of variable length or sites of expansion.

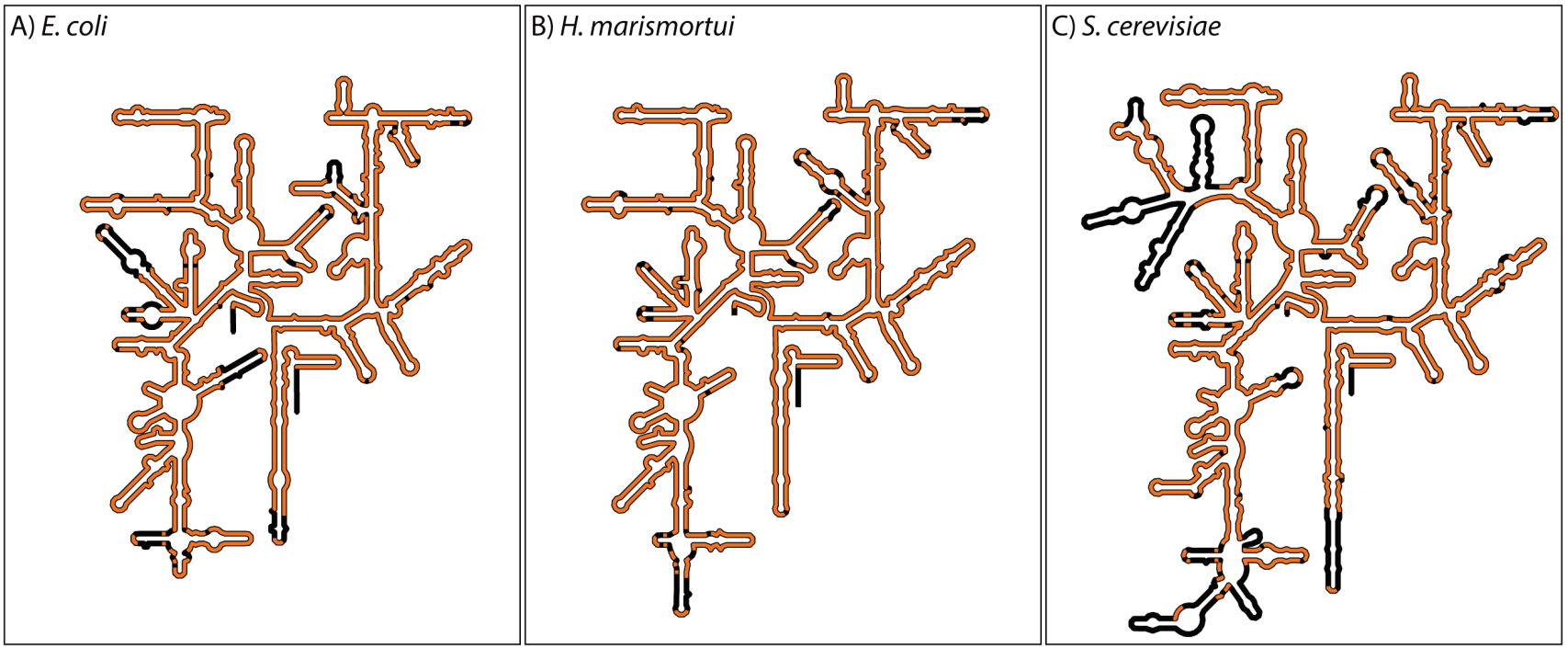

Figure 5.5. SSU rRNA secondary structures. A) *E. coli*, B) *H. marismortui,* and C) *S. cerevisiae.* The color indicates presence in the common core. Orange is included in the common core and black is excluded. Nucleotides are included if they are present in 95% of the species in the whole RiboZones alignment. Most black areas are helices of variable length or sites of expansion.

Most of the nucleotides in the prokaryotic ribosome (almost 90%) are in the prokaryotic common core. Domains II, IV, and V are almost entirely included. The majority of Domain I and VI, and about half of Domain III are included. All the major functional regions are present, including the PTC, sarcin/rincin loop, L7/L12 stalk, L11 binding domain, inter-subunit bridges, and the L1 binding domain.

The black areas are generally sites where helices are variable length or sites of expansion. A few isolated areas are domain or species specific insertions, un-alignable RNA loops, or variable length single stranded regions. Each region can be investigated, starting with the LSU. Helical labels are available on RiboVision and in our Ribosome Gallery, at http://apollo.chemistry.gatech.edu/RibosomeGallery/.

Helix 1 is black due to lack of sufficient sequence at the 5' and 3' ends of the gene. The gene ends are not always sequenced or annotated. Even when the surrounding sequence is available, the exact boundaries of the rRNA gene are ambiguous. rRNA does not have a start and stop codon like protein. In addition to sequence availability, eukaryotes do not have a Helix 1 and may have shorter ends.

Helix 7 is a hyper-variable region. The pattern of black nucleotides, along with analysis of the alignment data, shows that while all species have a Helix 7 of similar size, the shape and composition varies, so a perfect structure based alignment is difficult in this region. The perfect alignment would still exhibit many black nucleotides, because the number of base pairs and the number of single stranded nucleotides varies.

Helices 8, 9, and 10 are variable length. More than 5% of species have short helices in this region. Some species have only a tiny single stranded region instead of Helix 9.

Helices 16, 17, and 18 are variable length and structure. That makes a structure based alignment of those regions very difficult. All species have RNA sequence in this region, but it is difficult to define the boundaries between helices 16, helices 17, and helices 18 based on the alignment. Some species have short versions of these helices.

Helix 25 can be very short in certain bacteria species, up to extremely long in mammals. The minimum length of Helix 25 is shown. Most species have a longer Helix 25, with eukaryotes having a huge expansion segment here (ES7), which is apparent when looking at all three structures simultaneously. The tetra loop of Helix 25 is un-alignable due to the extreme variability of this region.

Helix 38 has multiple eukaryotic expansion site branching from it. Helix 38 includes single stranded and flexible regions. The alignment is ill defined around the black regions of Helix 38. Since Helix 38 hosts several expansion segments, the RNA alignment sometimes treats the edges of Helix 38 as part of the expansion segment. There is not a clear-cut boundary in the alignment of where exactly the common part of Helix 38 starts and the expansions begin. For most helices, the boundary is clear, but not in the case of Helix 38.

Helix 45 is very short in some species and expands greatly in other species. Therefore, the alignment does not contain a well-defined loop.

Helices 54, 55, 59, and 58 are highly variable both in sequence and in length. In addition, Helices 54 and 55 contain a site of expansion. While in general, most if not all species, have helices here, they are not in common. This region has extensively evolved since LUCA, and it is not known if LUCA had all these helices, or if they are all a product of convergent evolution.

Helix 63 is another helix that can either shrink or grow depending on species. The black helical region is indicative of helix shrinkage. The mostly black loop region is indicative of helix growth, when combined with the pattern from eukaryotes.

Helix 68 and 78 show a nicely aligned variable length helix. The tetra loop is well aligned and orange. As the eukaryotic structure (**Figure 5.4C**) shows, these helices shrink in eukaryotes, relative to at least some prokaryotes.

Helix 79 is variable length and a site of expansion and has a common core pattern consistent with this fact.

Helix 98 is optional. While most species have at least a small Helix 98, more than 5% lack a helix 98. Therefore, Helix 98 is not in the common core. It likely evolved by convergent evolution. Alternatively, many independent species have lost their Helix 98.

The small subunit is simpler than the large subunit but follows the same common core patterns. The 5' and 3' ends are black due to sequence availability. Helices 6, 10, 17, 26, and 44 are well-aligned helices that can shrink or grow. Helix 9 is a site of expansion, so can be either small or large, and difficult to align. Helices 16 and 39 have variable structure and length.

While these figures show the overall rRNA structure pattern clearly, they do not portray specific phylogenetic information. Visual figure analysis must be performed along with visual analysis of the alignment. However, visualizing and presenting a whole alignment is challenging. Only a small portion of an alignment can be visualized at a time.

Phylogenetic information can be visualized through analysis of modified versions of the common core. RiboZones software includes several optional data filters, and more

features can be added as needed. For example, the eukaryotic sequences in the RiboZones alignment were filtered out, and new common core figures generated (**Figure 5.6** and **Figure 5.7**). However, alignment filtering is not the same as making a new alignment. The presence of eukaryotes during the alignment building process may have slightly influenced the overall alignment, mostly in the immediate vicinity of expansion sites and some of the tetra loops. Nevertheless, alignment filtering eliminates some sources of gaps and makes better definitions of the common core. A prokaryotic-only common core better represents LUCA and is more suited for evolutionary studies.

We compare the three domain (orange) common cores to prokaryotes only (purple) common cores, to identify further differences between domains. There are improvements in the difficult to align regions, and improvements in the previously missing tetra loops, evidence that the alignment has the prokaryotic loop regions aligned well, only failing to align relative to eukaryotes. Helix 78 is purple in the prokaryotic structures and black in the three domain structures. Therefore, Helix 78 only shrinks in eukaryotes, not in bacteria or archaea. Analogous figures combining archaea with eukaryotes and excluding bacteria are also possible.

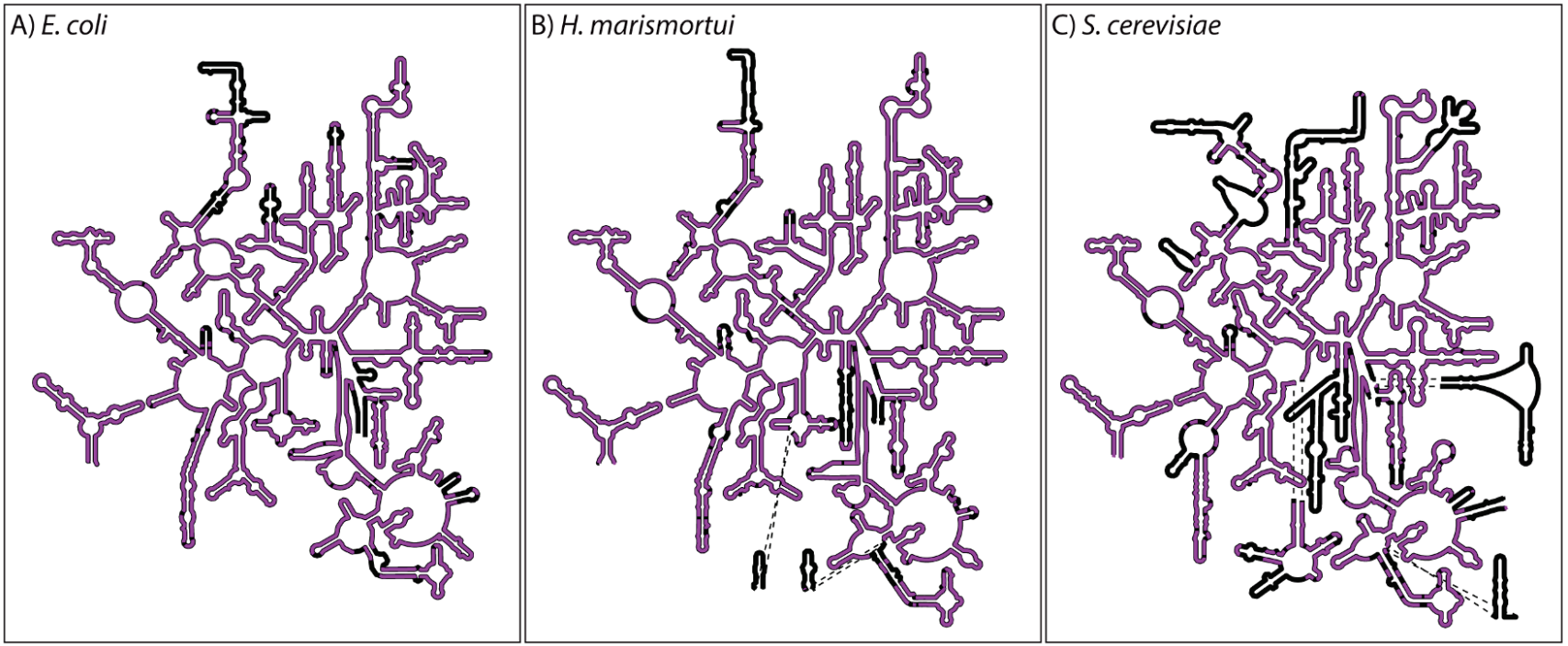

Figure 5.6. LSU rRNA secondary structures. A) *E. coli*, B) *H. marismortui,* and C) *S. cerevisiae*. The color indicates presence in the common core. Purple is included in the common core and black is excluded. Nucleotides are included if they are present in 95% of the *prokaryotic* species in RiboZones alignment. Only bacteria and archaea sequences are included. This is a better representation of LUCA. Most black areas are helices of variable length or sites of expansion.

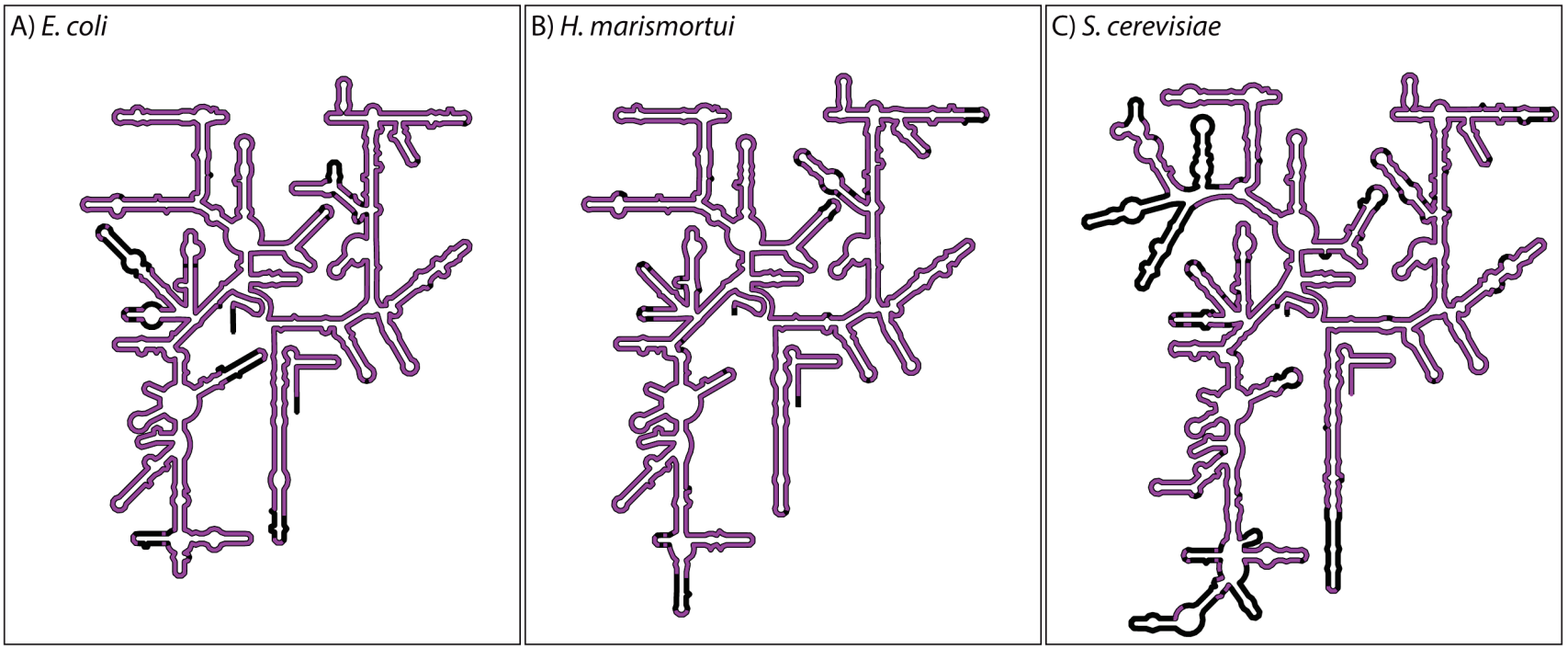

Figure 5.7. SSU rRNA secondary structures. A) *E. coli*, B) *H. marismortui,* and C) *S. cerevisiae.* The color indicates presence in the common core. Purple is included in the common core and black is excluded. Nucleotides are included if they are present in 95% of the *prokaryotic* species in RiboZones alignment. Only bacteria and archaea sequences are included. This is a better representation of LUCA. Most black areas are helices of variable length or sites of expansion.

Common core figures based on a single domain of life, in combination with the previous figures, allows additional information to be discovered, through strictly visual analysis. The differences between archaea and bacteria are isolated. In addition, the conservation of eukaryotes can be measured. **Figure 5.8** and **Figure 5.9** are using domain specific alignment filters, projected onto their respective representative structures. It is not technically necessary to match the alignment filters with their representative structures, if such figures are desired. RiboZones software can visualize, for example, an alignment of eukaryotes projected onto *E. coli.* For clarity, the bacterial alignment is represented in red, archaea in blue, and eukaryotes in green.

It is clear that Helices 9 and 98 can be eliminated in both bacteria and archaea. Without the domain specific figures, it would be necessary to examine the alignment to determine if helices 9 and 98 were optional in both domain. It also appears that Helix 98 can be eliminated in eukaryotes, but this is not true. The alignment of Helix 98 is not sufficient for eukaryotes.

It is now possible to determine that most archaea and most eukaryotes have Helices 15 and 30. Unfortunately, these figures are not sufficient to show that about half the bacterial species have a helix 15. There are no bacteria with a Helix 30, which is evident by looking at the alignment, but could not be known for sure with just common core figures.

For the SSU, it is now possible to see that Helix 9 is, at least 95% of the time, a longer version that only gets longer in eukaryotes. In contrast, Helix 10 can shrink in any of the domains of life. The unnamed helical insertion in *E. coli* Helix 33 is common to most bacteria; it is not specific to *E. coli*.

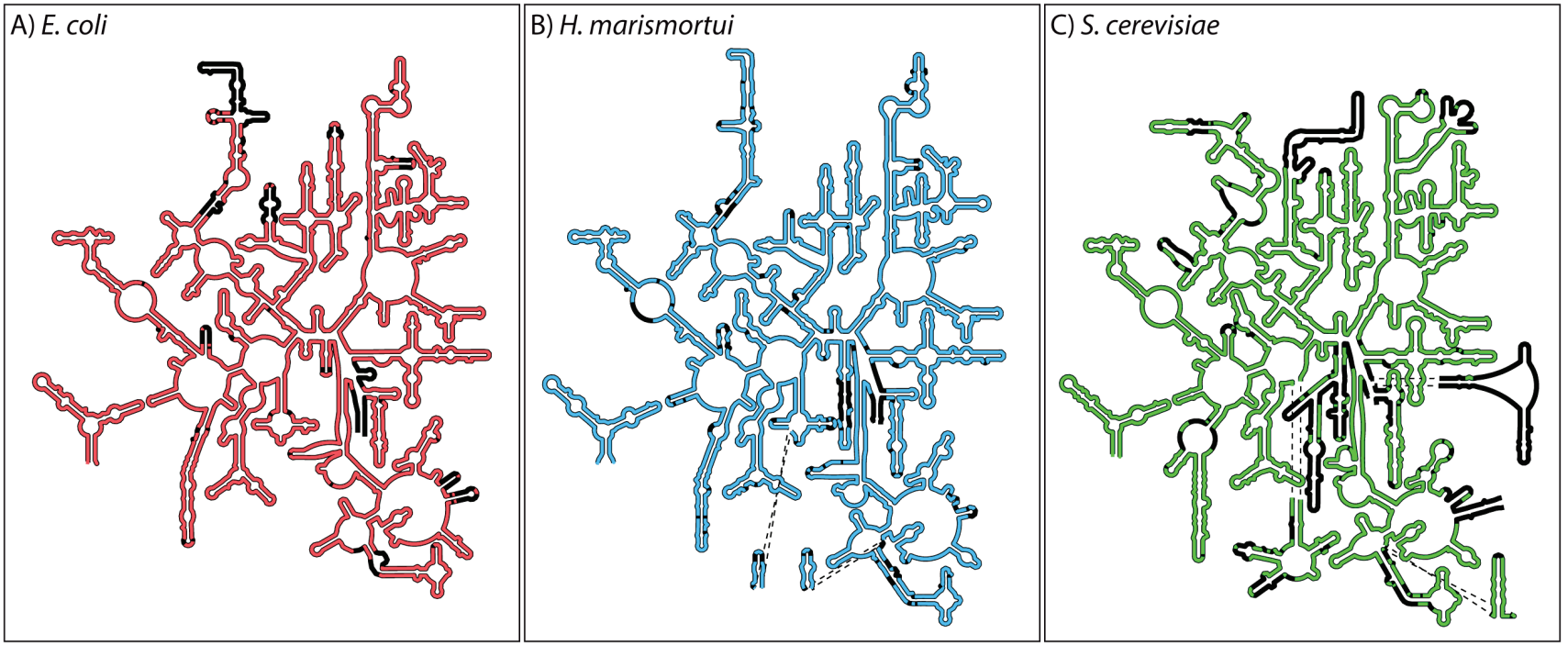

Figure 5.8. LSU rRNA secondary structures. A) *E. coli*, B) *H. marismortui,* and C) *S. cerevisiae.* The color indicates presence in the common core. Red, blue, or green are included in the common core and black is excluded. Nucleotides are included if they are present in 95% of the *domain specific* species in RiboZones alignment. Red counts just bacteria, blue counts just archaea, and green counts just eukaryotes. Note, these are subsets of the same alignment, not separate alignments.

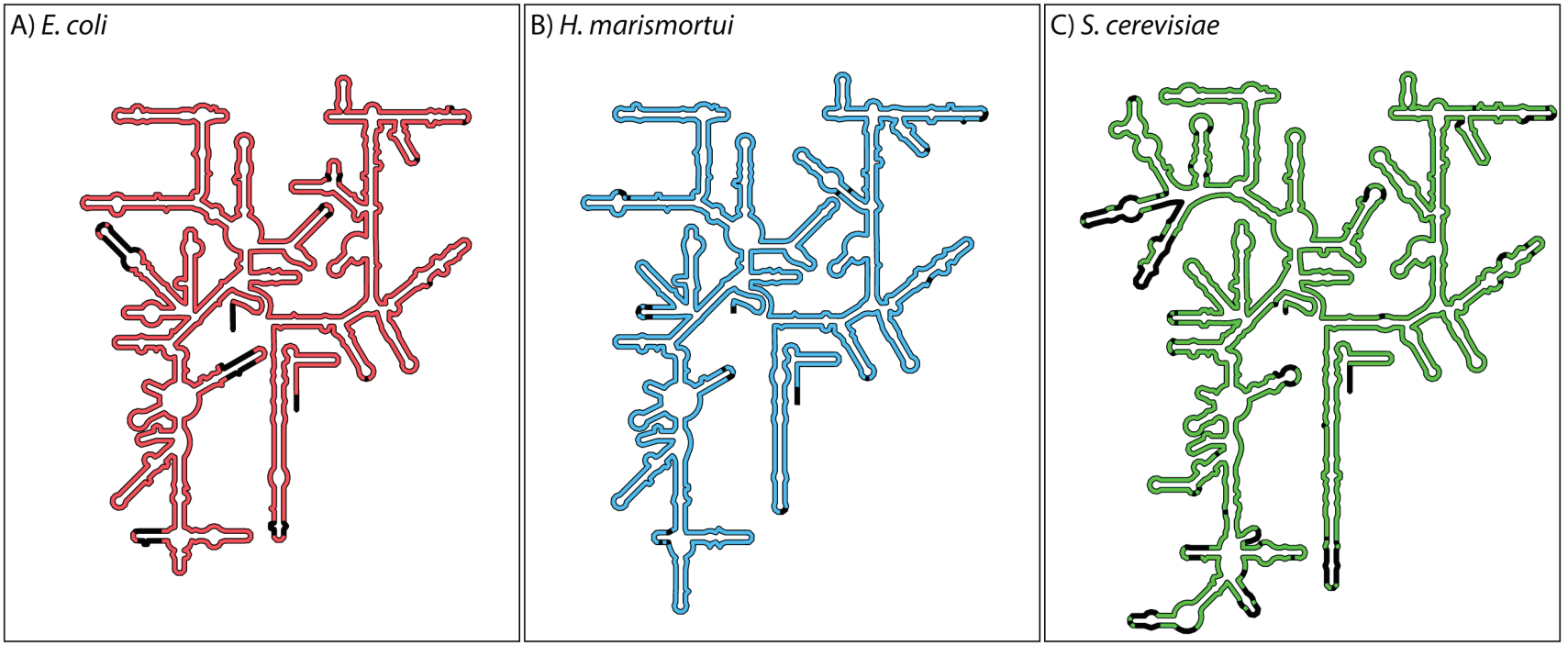

Figure 5.9. SSU rRNA secondary structures. A) *E. coli*, B) *H. marismortui,* and C) *S. cerevisiae.* The color indicates presence in the common core. Red, blue, or green are included in the common core and black is excluded. Nucleotides are included if they are present in 95% of the *domain specific* species in RiboZones alignment. Red counts just bacteria, blue counts just archaea, and green counts just eukaryotes. Note, these are subsets of the same alignment, not separate alignments.

RiboZones visualization techniques can be applied to the study of eukaryotes. Studying eukaryotes is more complicated than prokaryotes, but RiboZones is an invaluable tool. For eukaryotes, it is now clear that Helix 16 does not vary much amongst eukarya. The degree of conservation in the Helix 21 region can be estimated. Helix 33 is very conserved. In total, commonalties of all eukaryotes can be compared and contrasted with prokaryotes. Some parts of the ribosome are similar in bacteria and archaea, but different in eukaryotes. Other parts are similar in archaea and eukaryotes, but different in bacteria.

Deeper analysis requires data that are more detailed. For simplicity, and in the interest of more commonality, the focus shifts to the prokaryotic only version of the common core. RiboZones can calculate highly specific statistics in many combinations. Statistics should be matched to a specific hypothesis as too much detail just hinders understanding. Here, intermediate level data is presented.

The common core rRNA is classified into one of three classes 1) conserved base pairs, 2) conserved single nucleotides, and 3) non-conserved nucleotides. Conserved base pairs are defined the same way as in **CHAPTER 4**. Conserved single nucleotides are defined as those with a gap prorated Shannon entropy of greater than 0.56. This approximately corresponds to a frequency of 95%. Non-conserved nucleotides are all other common core nucleotides, those with less than 5% gaps, but not conserved according to the above definition.

The not common core rRNA, referred to here as divergent rRNA, is also classified into one of three classes 1) incomplete sequence data, 2) likely true divergent regions, and 3) unlikely true divergent regions. The 5’ end, 3’ end, and the 5S are put into

category 1. Some of the divergent rRNA is put into category 3, because it is likely to be in all prokaryotic ribosomes, but the sequence alignment is of insufficient quality in those regions. The differences between category 2 and 3 are subjective, further study needs to be done.

Figure 5.10 is a visualization of *E. coli* LSU rRNA. **Figure 5.11** is a visualization of *E. coli* SSU rRNA. **Figure 5.12** is a visualization of *H. marismortui* LSU rRNA. **Figure 5.13** is a visualization of *H. marismortui* SSU rRNA. Conserved base pairs are dark blue, conserved singles are medium blue, and non-conserved singles are light blue. Incomplete sequence data is gray. The likely divergent rRNA is dark red. The unlikely divergent rRNA is orange.

The divergent rRNA is mostly concentrated in a few areas in 3D. To better visualize the whole ribosome, a 3D model using both subunits of the ribosome assembled would be helpful. Since there is no 3D model for *H. marismortui*, only *E. coli* is presented. **Figure 5.14** contains the prokaryotic only version of the common core. Common Core rRNA is in purple, and divergent rRNA is in gray. There is also a tRNA molecule in yellow. The tRNA happens to be in the P/E hybrid state. A small piece of mRNA is in cyan.

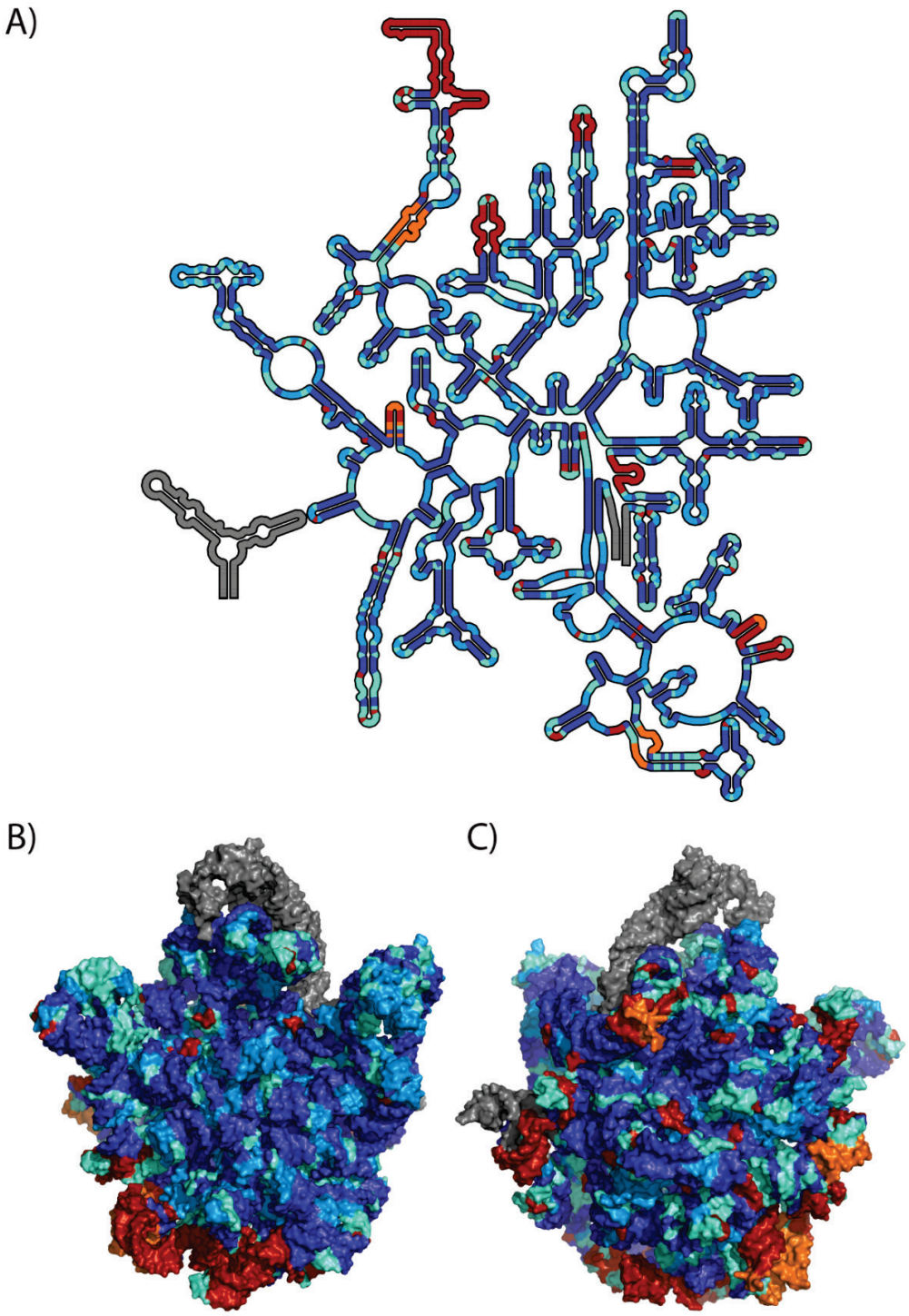

Figure 5.10. Prokaryotic Common Core for *E. coli* LSU rRNA. A) Secondary structure with detailed common core mapped onto it. Conserved base pairs are dark blue. Conserved single nucleotides are medium blue. Non-conserved single nucleotides are aquamarine. Divergent rRNA is red and orange. RNA without enough sequence data are gray. B) Same as in A, but on the 3D model of the LSU. C) Same as in B but rotated 180° around the y-axis.

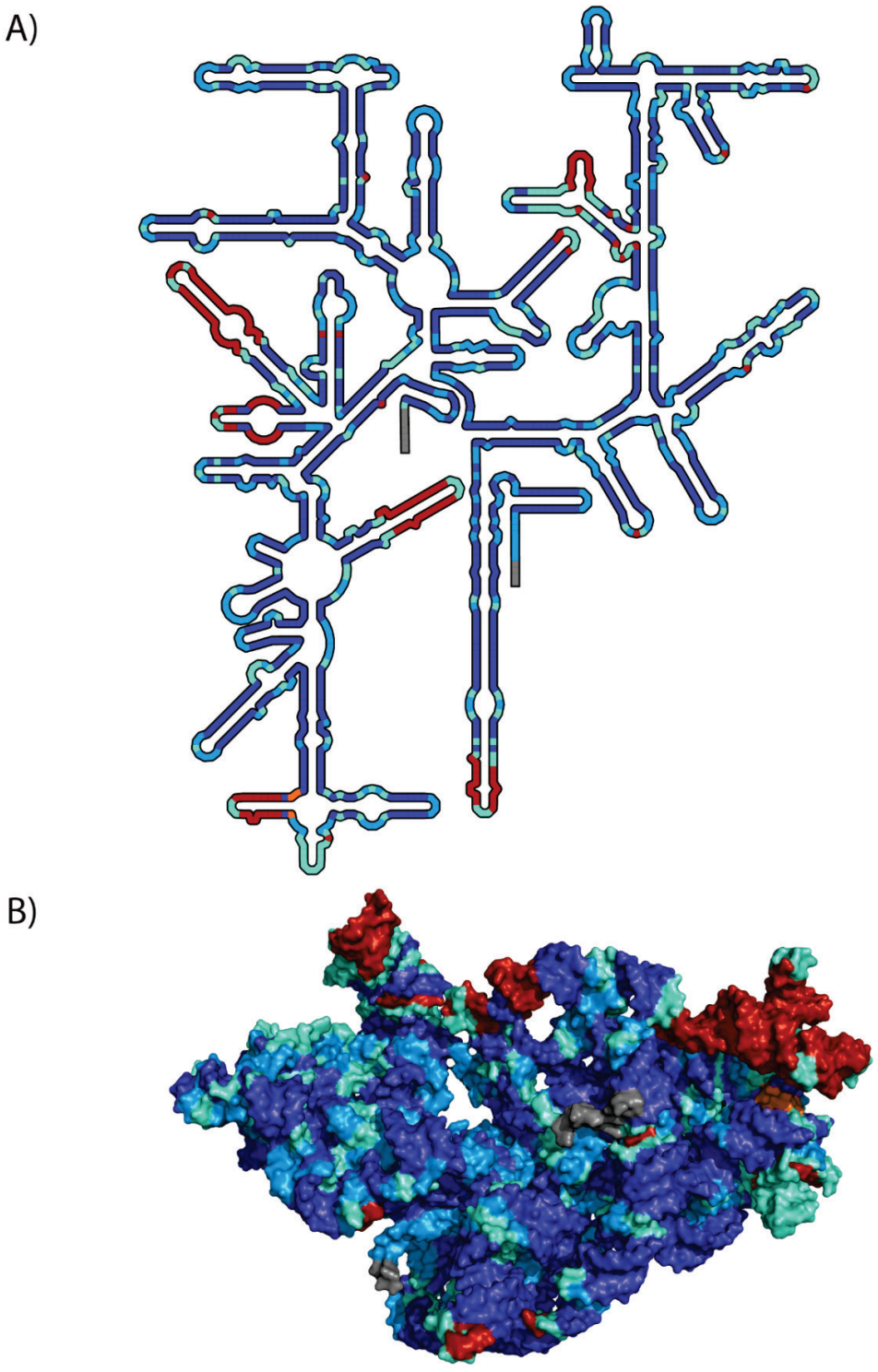

Figure 5.11. Prokaryotic Common Core for *E. coli* SSU rRNA. A) Secondary structure with detailed common core mapped onto it. Conserved base pairs are dark blue. Conserved single nucleotides are medium blue. Non-conserved single nucleotides are aquamarine. Divergent rRNA is red and orange. RNA without enough sequence data are gray. B) Same as in A, but on the 3D model of the SSU.

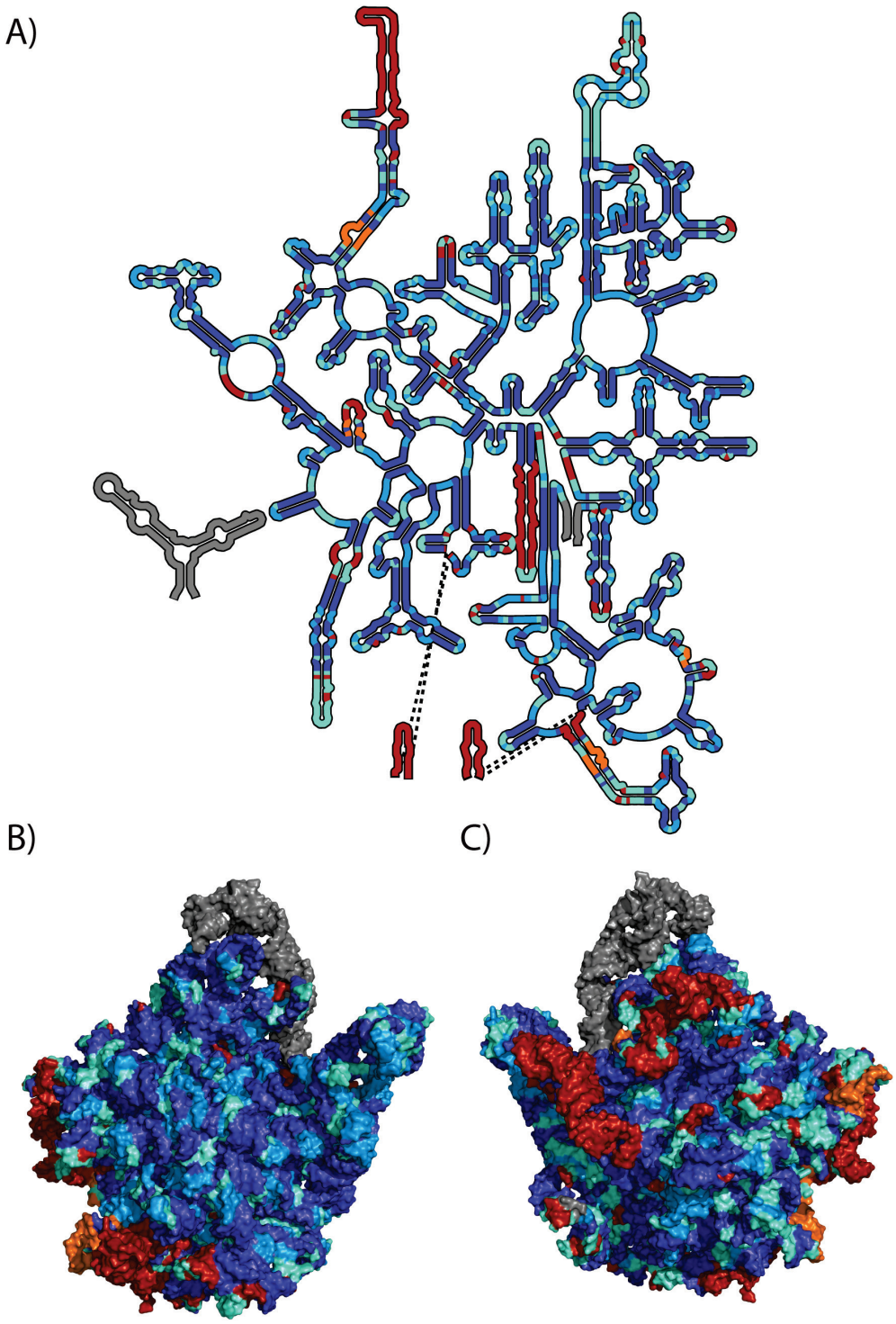

Figure 5.12. Prokaryotic Detailed Common Core for *H. marismortui* LSU rRNA. A) Secondary structure with detailed common core mapped onto it. Conserved base pairs are dark blue. Conserved single nucleotides are medium blue. Non-conserved single nucleotides are aquamarine. Divergent rRNA is red and orange. RNA without enough sequence data are gray. B) Same as in A, but on the 3D model of the LSU. C) Same as in B but rotated 180° around the y-axis.

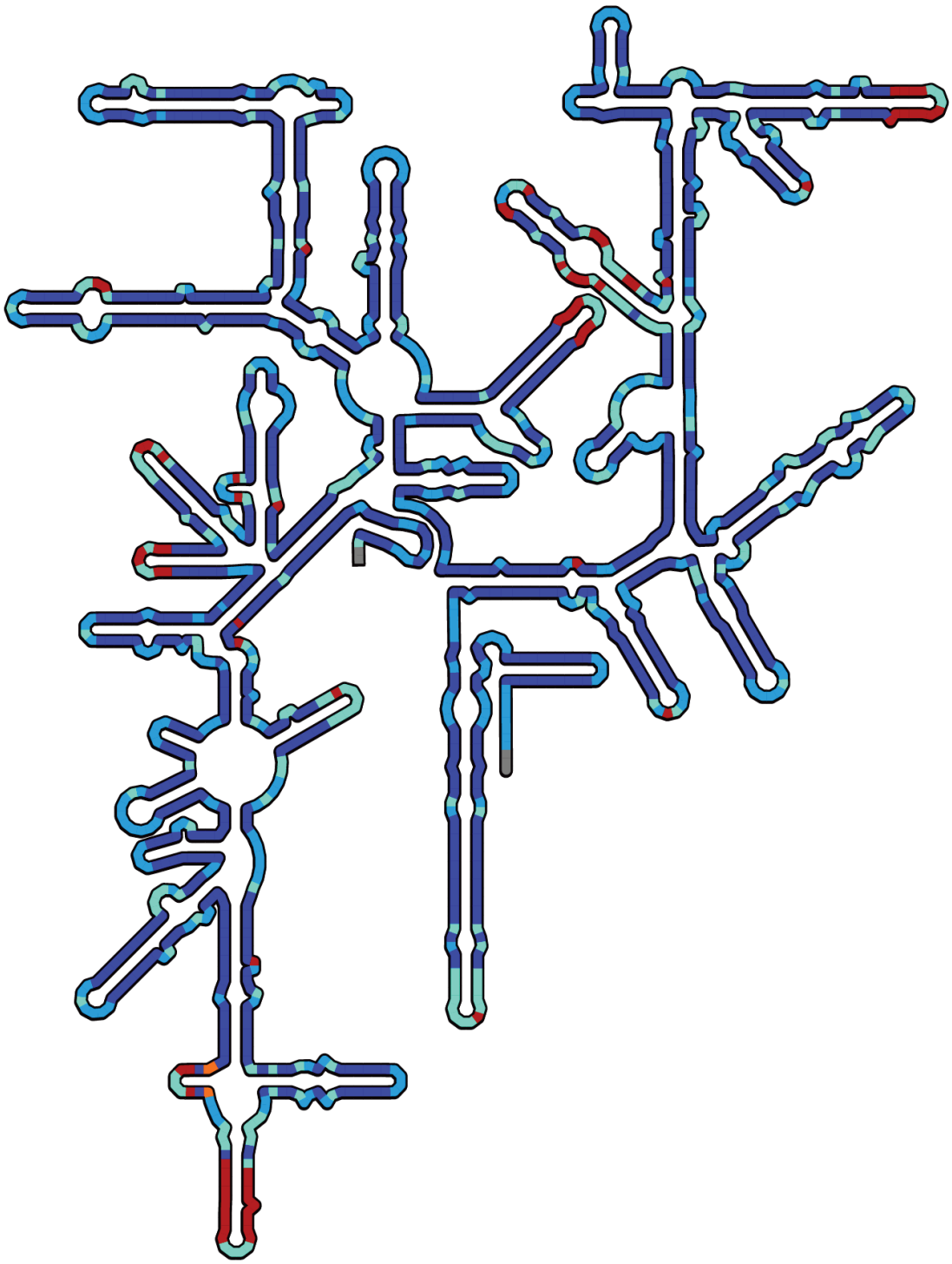

Figure 5.13. Prokaryotic Detailed Common Core for *H. marismortui* SSU rRNA. A) Secondary structure with detailed common core mapped onto it. Conserved base pairs are dark blue. Conserved single nucleotides are medium blue. Non-conserved single nucleotides are aquamarine. Divergent rRNA is red and orange. RNA without enough sequence data are gray.

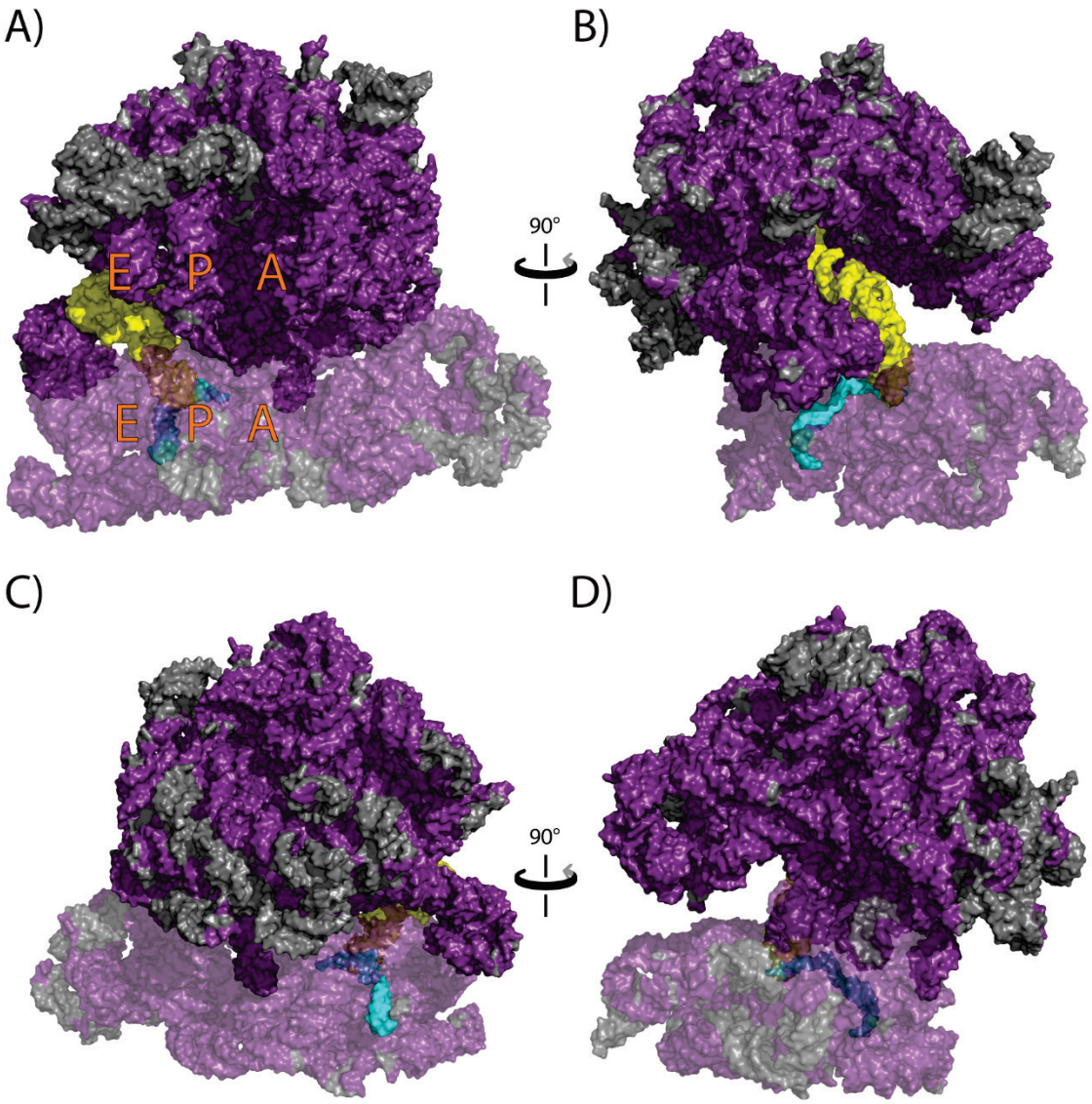

Figure 5.14. Prokaryotic Common Core for *E. coli* LSU and SSU rRNA. A) Common core rRNA is colored purple and divergent rRNA is colored in gray. A tRNA molecule in the P/E hybrid is shown in yellow. A short mRNA is shown in cyan. The approximate positions of the A, P, and E sites for both the LSU and the SSU are shown. View from the "front." In this orientation, the LSU is on top, and the SSU is on the bottom. The SSU is made partially transparent. The new tRNA molecules would enter from the right side. B) Same as in A, but rotated 90°around the y-axis.

5.5 Detailed Common Core Analysis

The common core as a model has significant scientific utility. Any independent parameter of the rRNA can be graphed, analyzed, and / or visualized with respect to the common core. Traditionally, most researchers only study one ribosome at a time. Analyzing and visualizing multiple ribosomes at a time leads to a greater understanding and higher scientific significance.

For example, average base pair adjusted Shannon entropy can be tabulated for multiple subunits, multiple subsets of RNA, and for multiple alignments (**Table 5.1**). Average base-pair adjusted entropy was calculated for both the LSU and SSU. The alignment was projected onto both *E. coli* and *H. marismortui*. Here, three alignments were used, the prokaryotic only alignment, the bacteria only alignment, and the archaeal only alignment. The RNA was divided into two subsets, common core, and divergent.

Table 5.1. Average base-pair adjusted entropy for several ribosomal subsets of common core vs divergent RNA.

	E. coli			
	Common Core [%]	Prokaryotic Entropy	Bacterial Entropy	Archaeal Entropy
LSU Common Core	88	0.33	0.25	0.29
SSU Common Core	90	0.26	0.19	0.2
LSU Divergent	12	1.69	1.61	1.73
SSU Divergent	10	1.68	1.64	1.87
	H. marismortui			
	Common Core [%]	Prokaryotic Entropy	Bacterial Entropy	Archaeal Entropy
LSU Common Core	88	0.36	0.28	0.31
SSU Common Core	94	0.24	0.18	0.18
LSU Divergent	12	1.76	1.78	1.53
SSU Divergent	6	1.76	1.83	1.5

For the LSU, both the bacterial and archaeal common core is around 88% of the total RNA of their respective representative species. For the SSU, the archaeal common core is slightly larger (94%) of the total RNA, than the bacterial common core (90%). Some of the helices in the bacterial SSU are allowed to get shorter and *E. coli* happens to be a relatively large representative for bacteria.

The conservation of the LSU is lower (higher entropy) than the conservation of the SSU in all cases. The influence of eukaryotes on this data is close to nonexistent, since no eukaryotic sequences are includes in analysis, nor are eukaryotic structures. Therefore, when studying the differences between bacteria and archaea, more effort should be focused on the LSU.

The bacterial LSU is slightly more conserved than the archaeal LSU, despite using a higher number of bacterial sequences. In contrast, the bacterial and archaeal SSU appear to be similarly conserved. For both subunits, the individual domain entropies are lower than the prokaryotic entropies. Therefore, there are certain regions of both the LSU and the SSU that are conserved within a domain of life, but different between them. RiboZones could be used to calculate and visualize those regions on a per-nucleotide level. That would be the first step in hypothesizing why those regions are different and what it means for ribosomal function and evolution.

Interestingly, the *E. coli* and *H. marismortui* statistics are very similar. This would be expected if the alignment was high quality and if the common core is genuinely common. While it is still a good idea to do theoretical experiments on multiple species, the fact that they are so similar can justify simplifications for figure preparation or for studying just one model species in wet lab experiments.

As an example of more detailed nucleotide level data portioned into common core and divergent RNA, fine-grained onion (**Figure 5.15** and **Figure 5.16**) is used. A fine-grained onion is the approximate distance of a particular residue from the functional core of a ribosomal subunit. Similar analysis techniques can be used for a wide variety of data.

Fine-grained onion data was mapped onto *E. coli* LSU and SSU rRNA. The prokaryotic common core was also mapped on these figures in the form of a black outline. Regions of rRNA without a black outline are not in the common core. RNA close to the PTC or the DCC is colored blue. Regions of rRNA far away from the PTC or DCC are colored in red. The common core RNA covers the whole color spectrum. The divergent RNA however, is heavily concentrated in yellow, orange, and red regions.

Table 5.2 contains aggregate statistics of the size of the common core. Including only the middle 95% of the nucleotides reveals the LSU and SSU common core to be of similar size, between 20 Å and 110 Å, with a mean of approximately 60 Å. The divergent RNA however has a range of about 50 Å up to almost 130 Å. The mean distance of the middle 95% of the divergent RNA is 89 Å for the LSU and 95 Å for the SSU. Despite the name, the SSU is actually longer than the LSU, in 3D, but its rod shape makes it overall smaller in volume.

Table 5.2. Aggregate statistics for fine-grained onion mapped onto the *E. coli* common core. The ranges and means of the nucleotide distance from the center of the subunits are shown.

	Common Core [%]	**Middle 95% Range [Å]**	**Middle 95% Mean [Å]**
LSU Common Core	88	22-106	63
SSU Common Core	90	21-111	62
LSU Divergent	12	53-110	89
SSU Divergent	10	51-129	95

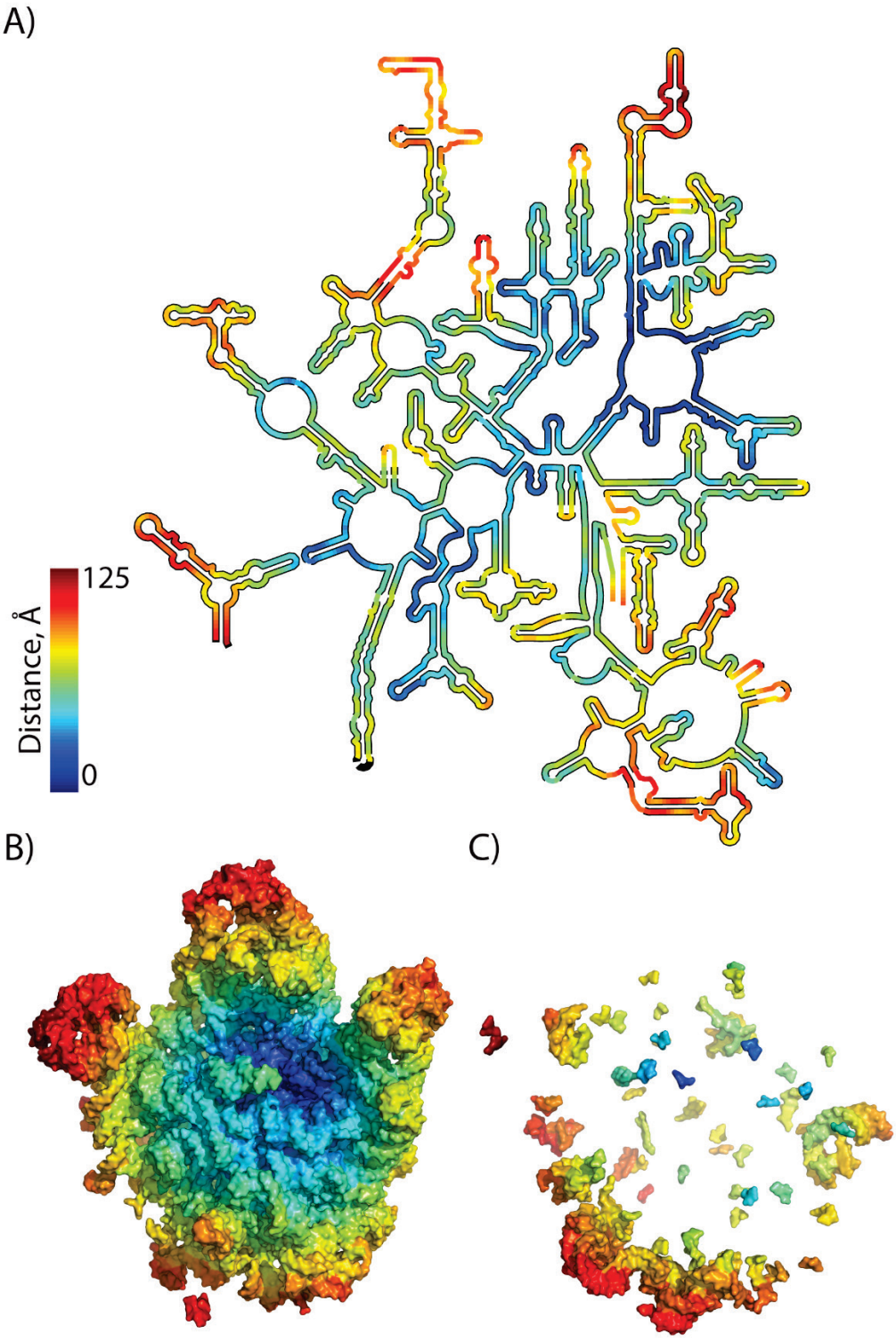

Figure 5.15. Fine-grained onion mapped onto *E. coli* LSU. A) Secondary structure of *E. coli* LSU RNA. Fine-grained onion is mapped as a colored contour line. The common core is outlined in black. B) 3D model of the common core with fine-grained onion. C) 3D model of the divergent RNA with fine-grained onion.

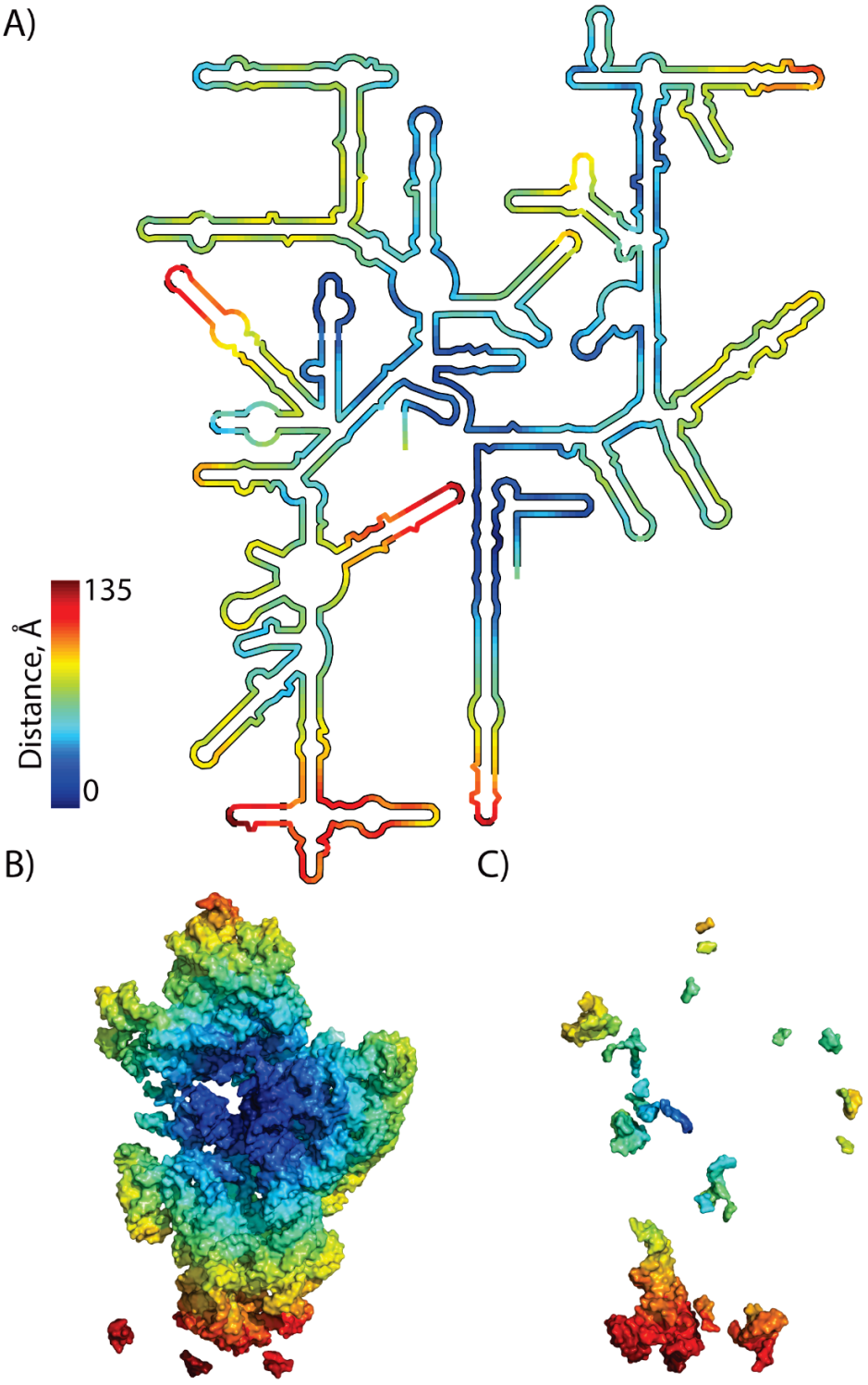

Figure 5.16. Fine-grained onion mapped onto *E. coli* SSU. A) Secondary structure of *E. coli* SSU RNA. Fine-grained onion is mapped as a colored contour line. The common core is outlined in black. B) 3D model of the common core with fine-grained onion. Free floating regions are the aligned RNA loops. C) 3D model of the divergent RNA with fine-grained onion.

5.6 Relatively recent eukaryotic expansions cause massive ribosomal growth

Bacteria and archaea rRNA is remarkably similar. Studying these molecules in detail would elucidate insight into structure, function, and the evolution of the ribosome. RiboZones has reduced the barrier of entry into the ribosome field and eliminated the previously rate-limiting step of quality figure production.

The main rRNA differences between bacteria and archaea are the lengthening of Helix 25 in archaea and the addition of Helix 15 and Helix 30 in archaea. About half the bacterial species have a Helix 15, but none has a Helix 30. It is clear that these helices grew onto an expanding common core.

Eukaryotic ribosomes continued to grow. **Figure 5.17** shows a series of whole ribosomes, from *E. coli*, to *S. cerevisiae*, to *H. sapiens*, including rProteins. Images are approximately on the same scale. Human ribosomes are about 50% larger than bacterial ones. Human has several naked rRNA helices sticking out in several directions. The role of these eukaryotic expansion segments is mostly unknown. Possibilities are that they play a role in regulation, initiation, membrane association, chaperone association, etc.

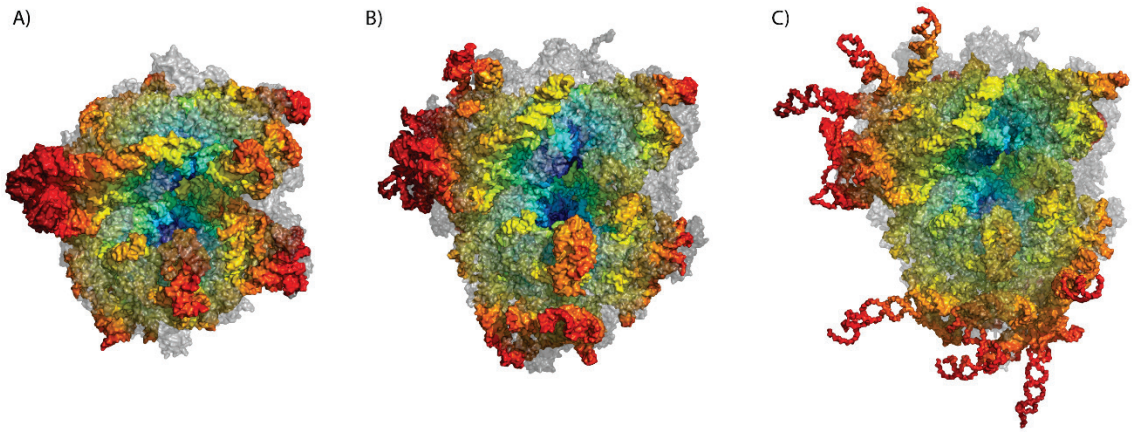

Figure 5.17. Fine-grained onion mapped onto whole ribosomes. Blue rRNA is close to the functional centers of the respective subunits, while red rRNA is remote. Ribosomal proteins are shown as transparent gray surfaces. A) *E. coli*. B) *S. cerevisiae, C) H. sapiens*.

Ribosomal size can be graphed as a function of time (**Figure 5.18**). Ribosomal structures themselves do not contain timing information. They are all modern ribosomes. However, they can be arranged into phylogenetic trees. For any given pair of species, the time since their common ancestor can be estimated through previous phylogenetic methods. For consistency, a single source of time estimates is preferred. The TimeTree book is used here. [113] The ribosomal size of the common ancestor is approximated as the minimal size of all ribosomes descended from that common ancestor.

Ribosomes have gone through several phases for growth. On the early prebiotic Earth, ribosomes had to evolve from simple RNA-like monomers up to the common core in only a few hundred millions years. While the time span and specific details are not known, that is an average rate of approximately five nucleotides per million years. It is likely that it actually grew a lot faster during shorter periods. Next, there was very little growth between bacteria and archaea for approximately one billions years. That is a rate of about 0.1 nucleotides per million years, two orders of magnitude smaller than on the prebiotic earth. The next phase is the development of basic eukaryotes. There was a growth rate of around 0.5 nucleotides per million years. The phase from the first multicellular organisms up to vertebrates is characterized by a growth rate of about one.

Mammal ribosomes are massively larger than all other ribosomes in the dataset. The growth rate from vertebrates to mammals is about four nucleotides per million years. The fastest stage of ribosomal evolution is the evolution of humans. Chimpanzees and humans diverged approximately 6 million years ago. Human ribosomes are 230 nucleotides larger. That is a growth rate of almost 40 nucleotides per million years.

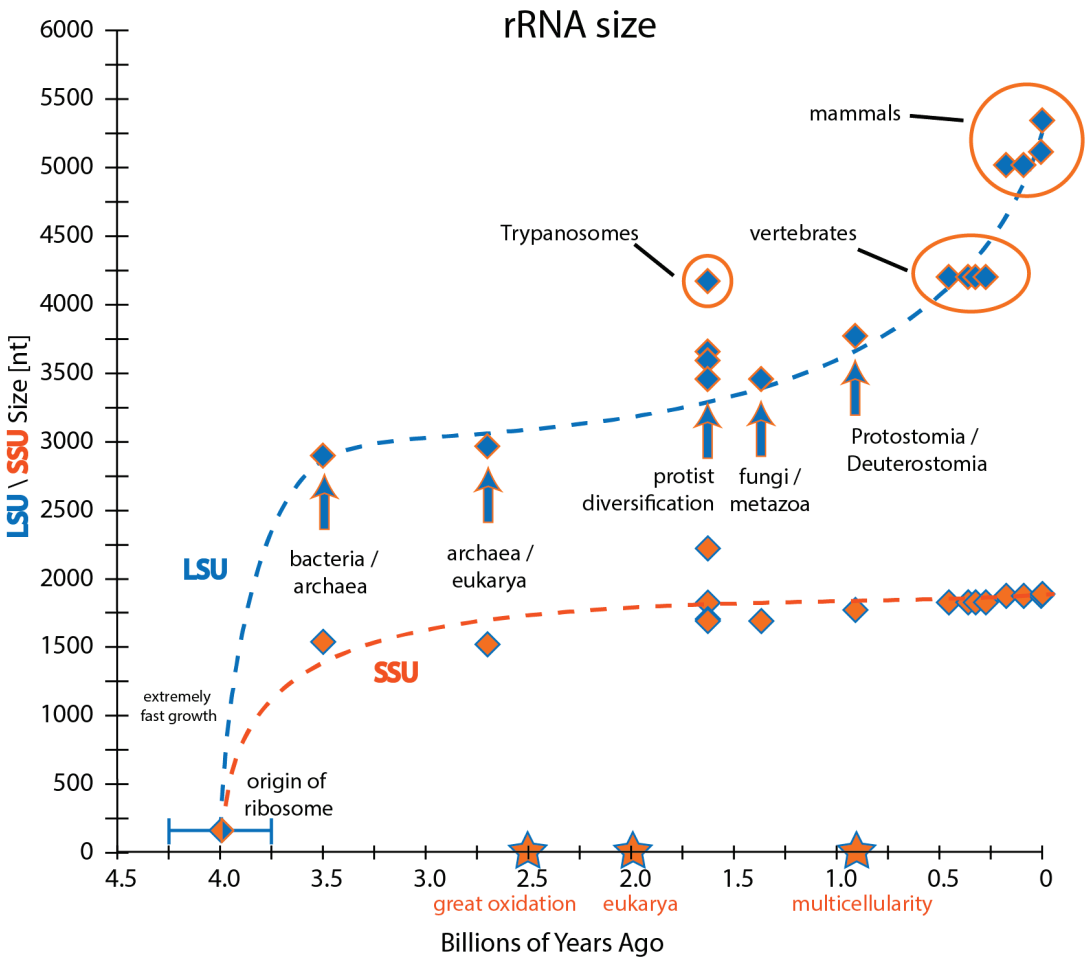

Figure 5.18. LSU and SSU rRNA size versus evolutionary time. Blue diamonds (LSU) and orange diamonds (SSU) represent points in time when common ancestors diverged in evolutionary history. The x-axis is time, in billions of years ago. Major events in the history of life on Earth are marked as orange stars. The y-axis is approximate size of the rRNA gene in the common ancestor of that point. The origin of the ribosome is approximately 4±0.25 billions of years ago. The dashed lines are rough estimates only.

The SSU has a similar, albeit overall slower, growth rate than the LSU. It appears that was an initial prebiotic period of fast growth. Then there was a long period of stasis. **Figure 5.18** actually shows a decrease, reflecting the fact that, at least in modern times, archaea SSUs are smaller than bacterial ones. Then there was an intermediate level of growth corresponding with the rise of eukaryotes. Once eukaryotes became multicellular,

the growth rate increased significantly for the LSU. However, the SSU does not have a sudden sharp growth rate in between vertebrates and the mammals, nor a large increase at human.

5.7 Discussion

The RiboZones structure-based alignment has been applied to build nucleotide level models of universal, prokaryotic, bacterial, archaean, and eukaryotic common cores. The concept of the common core has been discussed for decades, but statistically validated, nucleotide-resolution definitions, have never been attempted. A nucleotide level definition allows a variety of statistical parameters to be calculated, visualized, and analyzed. Simple entropy statistics and fine-grained onions are provided as examples of general-purpose data mapped onto a common core model.

A common core can be used for the basis of an evolutionary model. Each expansion segment can be analyzed for structural features and functional utility. One can address questions of where rRNA grows, how it grows, and why. These questions have two kinds of answers, both interesting.

At the physical level, there are various mechanisms of growth. The mechanisms would be especially different depending on phase of evolution being studied. Before there was DNA, RNA growth had to occur on the RNA and chemical level. The growth mechanism would include self-ligation of new RNA pieces. It is unknown to what extent templated synthesis existed, if at all. After the development of DNA, DNA level growth mechanisms take over. DNA genes can change due to errors, mutations, DNA polymerase slippage, etc. The “why” question is relatively uninteresting at the physical level. RNA grows because these chemical and biological mechanisms allow it to.

At the biological level, the why question is the most interesting. RNA growth should be conferring a genetic advantage. Each expansion is likely to serve a purpose. Ribosomal RNA is not only coevolving with ribosomal proteins, but also a vast translation and regulatory system. The availability of 3D molecular structures opens up a wide variety of both computational and physical experiments.

We do now know that the ribosome has evolved from the universal common core by accretion of new stable elements – the expansion segments. By comparing the 3D structures of the common core and eukaryotic organisms, we can now describe how the ribosome evolved from the common core to its modern structures observed in extant species. We hope to observe some regularities that govern the accretion process of eukaryotic expansions. We expect that the same accretion process was utilized by the ribosome during development of its common core. Our goal is to generalize these regularities and to propose an evolutionary model of the common core. The model is presented in **CHAPTER 6**.

CHAPTER 6:

A DETAILED PIECEWISE MODEL OF RIBOSOMAL EVOLUTION AT ATOMIC RESOLUTION

6.1 Introduction

The origin of life is the biggest unanswered question in biology. Darwin's Theory of Evolution explains the origin of the species, but says nothing about where the first species came from. In the early 20th century, it was still unclear exactly what "life" was. Did living things contain a special "life energy" that dead and nonliving things lacked? By the middle of the century, it was clear that living things were just complex sets of chemical reactions with certain properties such as self-replication and inheritance. The discovery of the structure of DNA, [115] the genetic code, [116-119] and the central dogma of Biology[120] pioneered the field of molecular biology. Applying the basic principles of evolution to biological molecules allowed the first modern theories of the origin of life.[121-123] It is clear that the translation system is an integral part of the transition from a nonliving world to a living world. [109,124-128]

6.1.1 Ribosomal Evolution Models

The origin of the translation system should obey the principles of evolution. Assuming some kind of RNA World, [117,119,129-131] the translation system should arise out of the Darwin Continuity Principle.[1] Evolution has no end goal, no foresight, and does not do complex things without reason. Evolution is a systematic process, and each step must make sense at the time. There generally has to be selective pressure for any step to survive. Wolf and Koonin present a detailed, yet cartooned, theoretical pathway for the evolution of translation.[1] Bernhardt and Tate present a similarly theoretical model on how

the transition from random peptide synthesis to coded protein synthesis could have happened.[132] These theoretical models do not contain molecular level details of what exactly the ribosome would look like.

Modeling the evolution of the ribosome started in the 2000's enabled by the publication of the first atomic-resolution 3D structures.[14,53,133] The first step was taken by Mears et al. in 2002, by modeling a minimal ribosome.[134] The minimal ribosome is too large to be the first step though. With the publication of *Thermus thermophilus* 70S ribosome structure in 2005,[135] Yonath et al. noticed that the ribosome's active site had pseudo symmetry. [110,136] They proposed that the proto-ribosome formed from the dimerization of two small RNA molecules that later evolved into the P-site and A-site. This provided a much-needed "first piece" of the ribosome. In 2009, Hsiao et al, proposed that the ribosome grows outward, like an onion, from the core PTC, composed primarily of the Yonath symmetry sites.

Analyzing intramolecular interactions helps build up more detailed models of ribosomal evolution. Fox et al. analyzed the interconnectedness of the traditional rRNA domains. [137] Smith et al. further analyzed molecular interactions including rProteins. They showed that the LSU PTC could fold on its own. It is likely that the SSU came later. Bokov and Steinberg[138] produced the most detailed model of LSU ribosomal evolution available, until the work presented here was published.[100] They divided the LSU into relatively simple pieces and modeled an evolutionary order which would be consistent with A-minor interaction information. However, the units themselves have no evolutionary or structural significance.

6.1.2 RiboZones Model

Here, we present the RiboZones model, the first atomic resolution model of both the LSU and SSU. The model breaks the rRNA into structural units with well-defined boundaries. The Darwin Continuity Principle is used to build up a rational model. Our model produces a relatively stable structure at each step, with ever-increasing functionality. Each stage in the model can be experimentally tested.

6.2 Modeling ribosomal growth

As shown in Chapter 5, the core structure of the ribosome never changes. Prokaryotic ribosomes are over 90% structurally conserved core with a few divergent expansions. Eukaryotic ribosomes are composed of the same common core, but with a larger number of expansions. The process of eukaryotic ribosome expansion should be analogous to the process of the expansion from the first proto-ribosome to the modern common core.

6.2.1 RNA growth over time

To understand the process of ribosomal growth over evolutionary time, a phylogenic tree of ribosomal structures was made. Helix 25 / ES 7 of the LSU is used as an example (**Figure 6.1**). By comparing the three-dimensional structures over evolutionary time, the steps in rRNA expansion can be ‘observed. This approach incorporates an assumption that the common ancestor of a pair of ribosomes is best approximated by the subset of rRNA that is present in both species. This subset of rRNA is, typically, most similar to the smaller rRNA. The general pattern is that as eukaryotic organisms increase in overall complexity, the rRNA becomes longer. However, during

evolution of individual species, some expansion segments decrease in size. The reduction of rRNA also occurs during the evolution of some archaeal or bacterial species. [100]

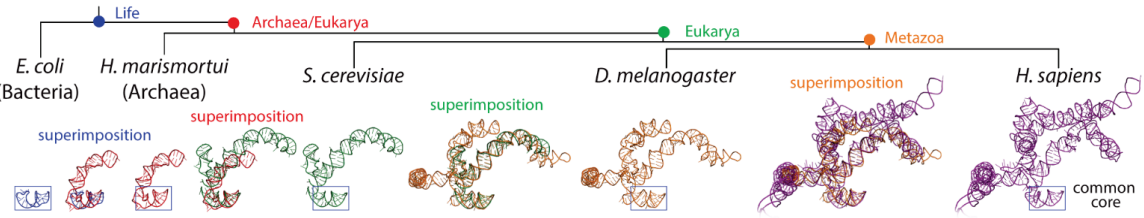

Figure 6.1. The evolution of Helix 25 / ES 7 shows serial accretion of rRNA onto a frozen core. This image illustrates at the atomic level how Helix 25 of the LSU rRNA grew from a small stem loop in the common core into a large rRNA domain in metazoans. Each accretion step adds to the previous rRNA core but leaves the core unaltered. Common ancestors, as defined in Figure 5.1, are indicated. Pairs of structures are superimposed to illustrate the differences, and to demonstrate how new rRNA accretes with preservation of the ancestral core rRNA. Each structure is experimentally determined by x-ray diffraction or Cryo-EM.[100]

A 'movie' of rRNA growth is exemplified by the lineage of expansion segment 7 (ES 7), as shown in Figure 6.1. A stem-loop of rRNA (Helix 25) and its progeny rRNAs present a multistep model of evolution of an rRNA domain (ES 7), at high resolution in three dimensions. ES 7 originates with a short 22-nucleotide stem-loop in the last universal common ancestor, which is approximated here by *E. coli*. This stem-loop grows to an 80 nucleotide bent helix in the common ancestor of Archaea and Eukarya. The common ancestor of Archaea and Eukarya is approximated by the archaeon *H. marismortui*. In the next step, ES 7 grows to a branched 210-nucleotide structure in the common ancestor of eukaryotes, which is approximated by *S. cerevisiae*. In the next step, ES 7 grows to a 342-nucleotide structure in the common ancestor of metazoans

(approximated by the arthropod *D. melanogaster*). Mammalian rRNA grows further, exemplified by the 876-nucleotide ES 7 domain in *H. sapiens*. [100]

In this series, one can observe accretion at the atomic level. The foundational Helix 25 stays intact in all larger rRNAs (**Figure 6.1**) and remains structurally conserved during a long evolutionary process. In general, each expansion step builds on preexisting rRNA, without substantially perturbing its 3D structure. This process has consistently been ongoing as the rRNA nearly doubled in size over 3.5 billion years of evolution, using the prokaryotic LSU as a foundation for the massive metazoan LSU.[100]

6.2.2 Insertion Fingerprints

The available structures allow us to make direct comparisons of pre- and post-expanded rRNA, and to observe rRNA conformation at sites where expansion elements join common core rRNA. We call the patterns observed at these sites *insertion fingerprints*. [100]

The predominant insertion fingerprint is a helical trunk joined to a secondary branching helix at a highly localized three or four-way junction[139] that minimally perturbs the trunk helix. At most, a few base pairs of the trunk rRNA is disrupted or unstacked at the site of insertion. These atomic-level fingerprints are seen by comparing many pre- and post-inserted expansion sites. For example, Helix 52 (**Figure 6.2A**) and Helix 38 (**Figure 6.2B**) are common core trunks in *E. coli* that have grown branches in the rRNA of *S. cerevisiae*. The *E. coli* rRNA shows trunk Helices 38 and 52 before insertion of the branching helices, while the *S. cerevisiae* rRNA shows trunk helices sporting branch helices after insertion. A second type of expansion is elongation of a previous helix. Helix 101 of *E. coli* is elongated in *S. cerevisiae* (**Figure 6.2C**), to form a continuous

stack within the previous helical element. Helix elongations do not leave distinctive structural fingerprints. Comparisons of pre- and post-expanded rRNAs reveal that helix insertions or elongations occurred within the common core in Helices 25, 30, 38 52, 54, 63, 79, 98, and 101 of the LSU rRNA.[100]

The patterns of conformation at sites of rRNA expansion suggest the reverse process, which is excision of inserted helices followed by re-ligation to generate the ancestral RNA (**Figure 6.3**). The expansion is predicted to be conformationally facile and readily reversible *in silico*. In general, a branching helix at an insertion fingerprint can be computationally excised, and the trunk rRNA can be re-ligated by subtle shifts in the positions of a few nucleotides or even a single phosphate group. The re-ligation can be achieved by gentle energy minimization with a shift of local atomic positions by a few Ångströms and minimal perturbation of the trunk rRNA. Our modeling demonstrates how rRNA can be expanded (and contracted) with preservation of the ancestral core.[100]

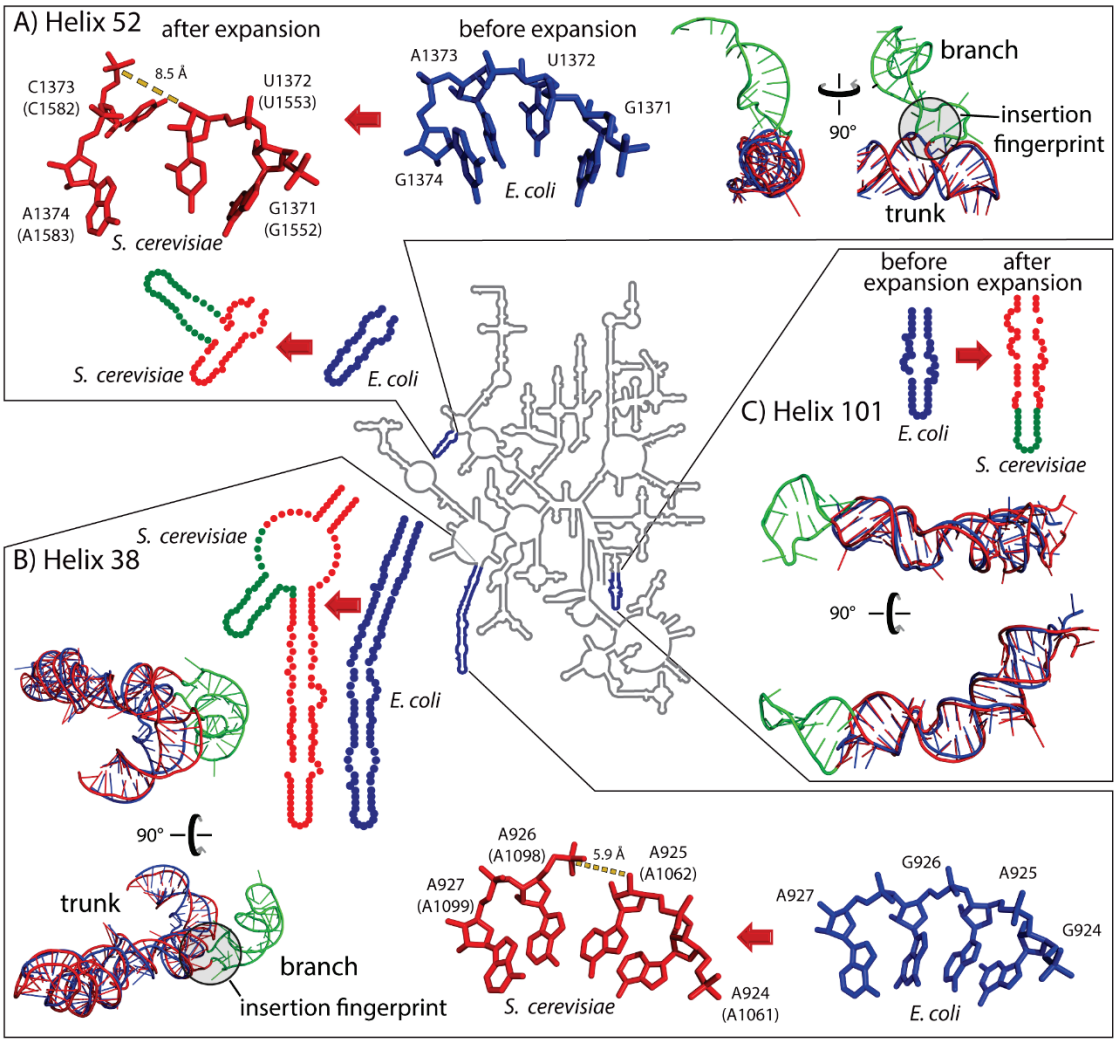

Figure 6.2. rRNA expansion elements in two and three-dimensions. A) Helix 52 is expanded by insertion. B) Helix 38 is expanded by insertion. C) Helix 101 is expanded by elongation. The secondary structure of the LSU common core rRNA, represented by that of *E. coli* (34) is a gray line at the center of the figure. Selected regions where the *E. coli* rRNA has been expanded to give the *S. cerevisiae* rRNA are enlarged. In the enlargements, the rRNA is blue for *E. coli* and red for *S. cerevisiae*, except that expansion elements of *S. cerevisiae* rRNA are green. These 'observed' expansion processes, from blue rRNA to red/green rRNA, are symbolized by red arrows. Superimposed pre-and post-expanded rRNAs indicate trunk (old) and branch (new) elements. Insertion fingerprints, where trunk meets branch, are highlighted by gray circles. *E. coli* nucleotide numbers are provided, with *S. cerevisiae* numbering in parentheses.[100]

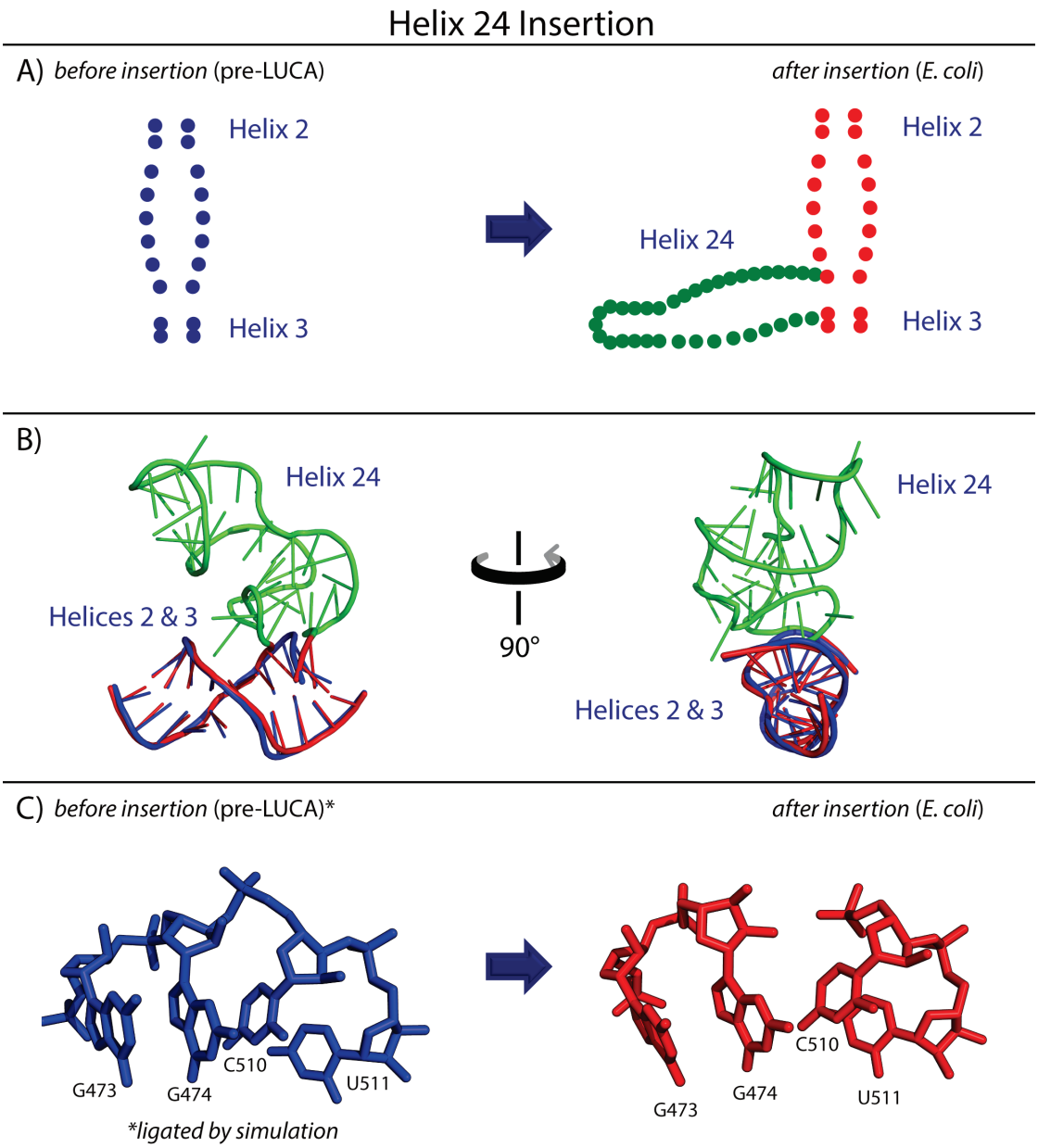

Figure 6.3. Expansion by helix insertion in the common rRNA core. Helices 2-3 (trunk) are expanded by insertion of Helix 24 (branch). A) Secondary structures of the trunk and branch fragments. B) 3D structures of the trunk and branch fragments. C) Atomic resolution representation of the insertion site. The pre-insertion state (blue) was modeled by computational ligation. Inserted branch is green and post-inserted trunk is red. The insertion process, moving forwards in time, is symbolized by blue arrows.[100]

6.3 Partitioning into ancestral expansion segments

Insertion fingerprints allow stepping back in time, to reconstruct the growth of the common core rRNA. The LSU and the SSU can be partitioned into smaller segments of rRNA bounded by insertion fingerprints that are named *ancestral expansion segments* (AES's). Insertion fingerprints are observed deeply buried within the common core of the LSU and the SSU. These ancestral insertion fingerprints appear identical in form to modern insertion fingerprints of eukaryotic expansions. The observation of ancestral insertion fingerprints suggests that addition of eukaryotic expansion segments followed patterns established in biological antiquity. The ancestral insertion fingerprints within the common core point to some of the oldest imaginable evolutionary events, and imply a method to work backwards in time, to identify pathways of expansion during formation of the common core.

The LSU and the SSU have been broken up into AES's (**Figure 6.4** and **Figure 6.5**). The LSU is about 70 AES's and the SSU is about 40 AES's. Each AES is delineated by an insertion fingerprint, although some fingerprints are no longer clear due to multiple insertions and distortions. AES's are both structural and evolutionary units. Some AES's correspond with traditional helices, but many do not. Many AES's are two or more traditional helices plus single stranded regions.

It is not possible to define AES's based on secondary structure; they are strictly a 3D phenomenon. In 3D, each AES looks like a distorted helix, with a single, possibly heavily curved, helical axis. There are a numerous three-way and 4-way junctions. Traditionally, a 4-way junction would be considered 4 independent helices, but in 3D, they can look like two helices fused in their centers.

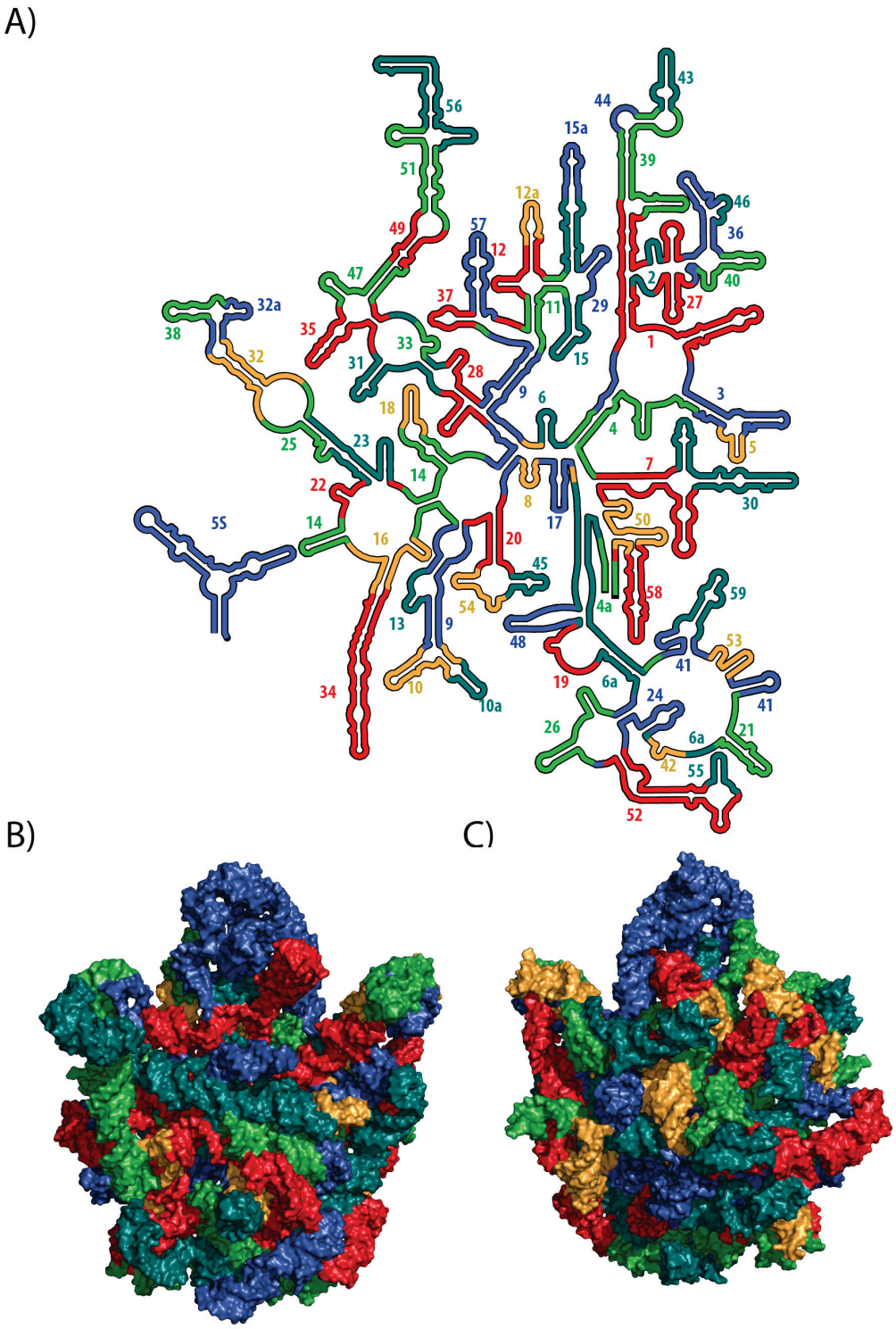

Figure 6.4. rRNA evolution mapped onto the LSU rRNA secondary structure of *E. coli*. The common core is built up in phases, by stepwise addition of ancestral expansion segments (AESs) at sites marked by insertion fingerprints. Each AES is individually colored and labeled by temporal number. AES colors are arbitrary, chosen to distinguish the expansions, such that no AES is the color of its neighbor.

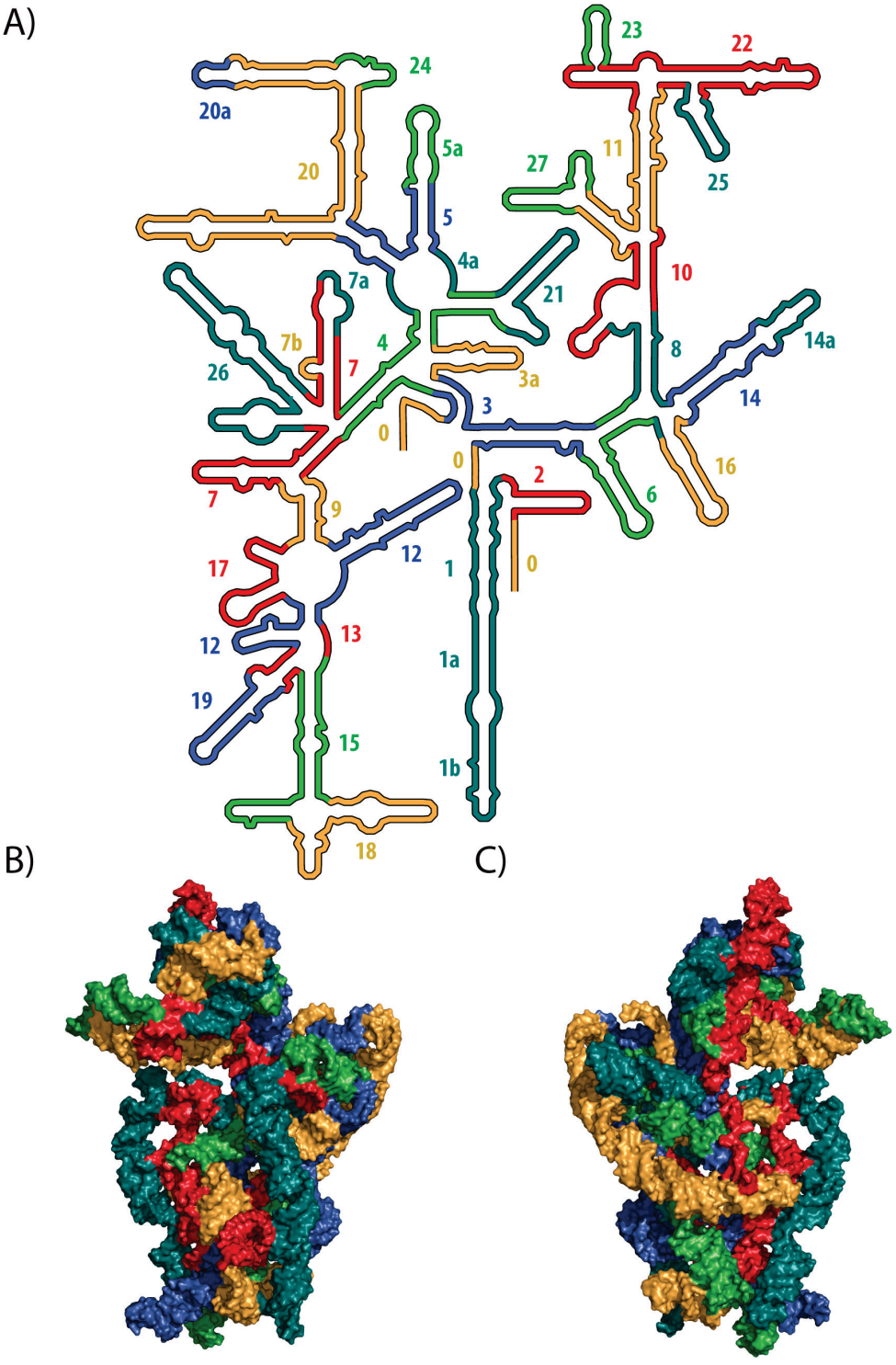

Figure 6.5. rRNA evolution mapped onto the SSU rRNA secondary structure of *E. coli*. The common core is built up in phases, by stepwise addition of ancestral expansion segments (AESs) at sites marked by insertion fingerprints. Each AES is individually colored and labeled by temporal number. AES colors are arbitrary, chosen to distinguish the expansions, such that no AES is the color of its neighbor.

Secondary structures of rRNA are misleading. Each structure has many base pairs that are not being shown on the secondary structure. Some base pairs are impossible to represent in the 2D plane of a secondary structure, except through confusing circuit diagram lines. Other base pairs could theoretically be represented accurately, but would require extensive modifications of the shape of the secondary structure. The current shape has been around for over 30 years, it might take many years for the community to reach a consensus on a new shape. RiboZones bridges the gap between scientists who prefer secondary structures and those who prefer tertiary structures, so that the optimal representation can be used at all times.

6.4 Separate Evolutionary Model for the LSU and SSU rRNA

A model of rRNA evolution based on AES's is evolutionarily grounded. If AES's are the evolutionary unit, rather than helices and domains, then the structural core of the ribosome is preserved as much as possible throughout the process. Each stage of evolution, having an intermediate ribosome, is structurally stable, consistent with the Darwin continuity principle. An AES-based model is also consistent with pre-DNA world RNA synthesis. In the earliest stages of molecular evolution, the RNA fragments could have been synthesized randomly and self-ligated together to form a stable proto-ribosome.

An AES-based model is convenient for *in vitro* and *in vivo* experiments. AES boundaries are natural places to cut out segments of rRNA to make a smaller ribosome, either for purposes of model reconstruction or studying the role of the excised segment. In most cases, producing a single strand of RNA/DNA for experiments would require little to no changes in sequence. At AES boundaries, it is often sufficient to just let the

rRNA ligate at that position naturally, alternatively needing only one additional nucleotide for spacing in a few cases. Dividing the ribosome outside of AES boundaries often requires adding an unnatural helix and tetraloop, which could interfere with the experiment.

A complete evolutionary model would be the ordering of the appearance of each AES in chronological order. However, the ribosome evolves outward from its center in multiple directions simultaneously. A complete linearization may not be possible, as some AES's may appear relatively simultaneously. One possible ordering is implied by the numbering scheme used in **Figure 6.4** and **Figure 6.5**. Future studies, both computational and experimental, will almost certainly change the ordering of at least some of the AES's. Nevertheless, our model is a good starting point and sets up a framework for producing alternative models.

A unique feature of the RiboZones model is breaking the peptidyl transferase center (PTC) into its components. The ribosome needs a first piece, AES1. The ancestor of AES1 could have been a ribozyme in the primitive, RNA-based world, possible binding to primitive amino acids, or similar prebiotic molecules. It may have catalyzed polyester formation, for example. We describe the origin of the PTC, the first few steps in our evolutionary model.[100]

The PTC is an essential component of the ribosome, responsible for peptide bond formation. The PTC is thought to predate coded protein[109,140] and is believed to be among the oldest polymeric elements of biological systems. The rRNA that forms the PTC (**Figure 6.6**) contains four insertion fingerprints. A single continuous trunk helix (red) with a defect at the base of the P-region appears to be the ultimate ancestor of the PTC.

This rRNA fragment, denoted as AES 1 (Ancestral Expansion Segment 1), is joined by AES 2 (the P-loop) at an insertion fingerprint (**Figure 6.6A**). AES 1 and AES 2 together comprise the P-region. AES 1 is also joined by AES 3 at a second insertion fingerprint (**Figure 6.6B**). The temporal ordering of the additions of AES 2 and AES 3 to AES 1 is undetermined.

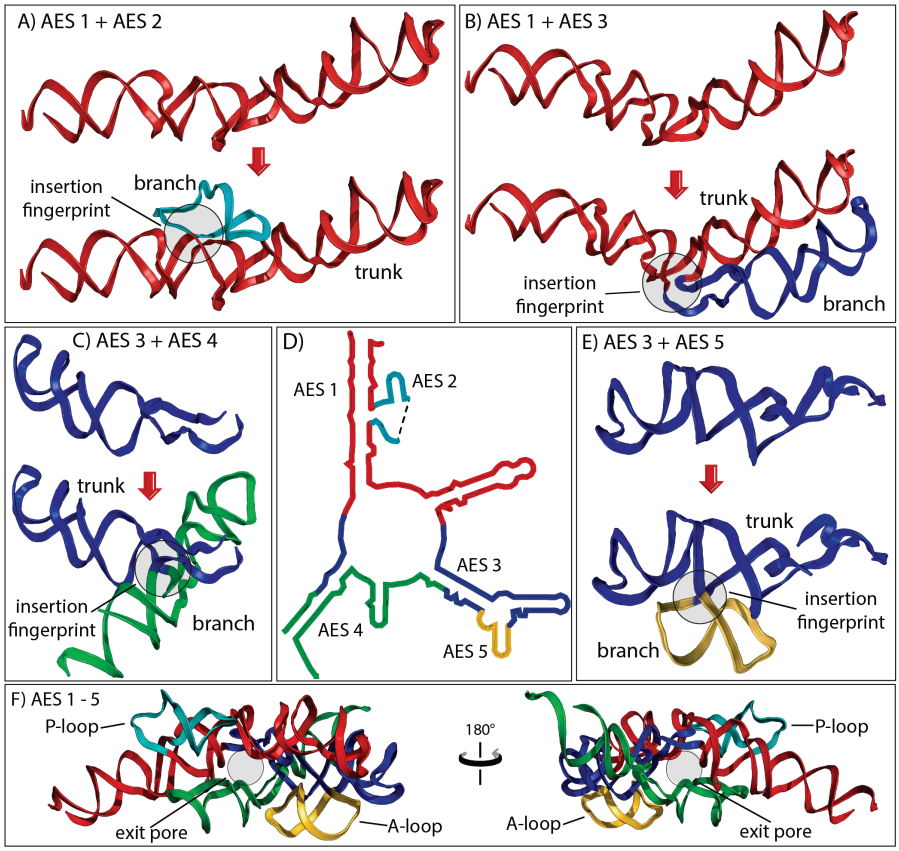

Figure 6.6. Origins and Evolution of the PTC. Trunk rRNA is shown *before* and *after* insertion of branch helix. A) AES 1 (red) is expanded by insertion of AES 2 (teal). B) AES 1 is expanded by insertion of AES 3 (blue). C) AES 3 is expanded by insertion of AES 4 (green). D) The secondary structure of AES's 1-5, which form the PTC and the exit pore (Helices 74, 80, 89, 90, 91, 92, and 93). The ends of AES 2 are located in direct proximity to each other in three-dimensions, indicated by a dashed line in the secondary structure. E) AES 3 is expanded by insertion of AES 5 (gold). F) The three-dimensional structure of AES 1-5, colored as in panels A-E. In each case, the *before* state was computationally modeled by removing the branch helix and sealing the trunk using energy minimization protocols. Positions of the P-loop, the A-loop, and the exit pore are marked.

AES 2 appears to be expanded in turn by the addition of AES 4 (**Figure 6.6C**) at one insertion fingerprint and the addition of AES 5 (the A-loop, **Figure 6.6E**) at a second insertion fingerprint. AES 3-5 form the A-region of the PTC and the Exit Pore, which is the entrance to the exit tunnel. By the method of Steinberg [138], AES 4-5 appear to be added after AES 2-3. In our model, AES 1 and four expansion segments (AES 2-AES 5) together form not only the A- and P-regions but also a pore that, with later expansions, develops into the exit tunnel (**Figure 6.6F**). In sum, we have a well-grounded model for evolution of some of the oldest polymeric elements in all of biology.

The approach described here is readily extended, leading to a stepwise model of evolution of the common core and beyond (**Figure 6.4**). We propose that functional elements of the LSU emerge in a specific progressive ordering, in a series of distinct phases (**Figure 6.7**).

Phase 1) Folding, Rudimentary Binding, and Catalysis. AES 1-2 is a branched duplex with a defect that forms the P-loop. This defect may confer catalytic activity[1] and/or ability to bind specifically to small molecules.

Phase 2) Maturation of the PTC and Formation of an Exit Pore. Inclusion of AES 3-5 adds the A-region to the P-region, in concert with formation of an exit pore.[141]

Phase 3) *Early Tunnel Extension.* Inclusion of AES 6-10 extends the exit pore, creating a short tunnel. The stability and rigidity of the tunnel are increased by buttressing.

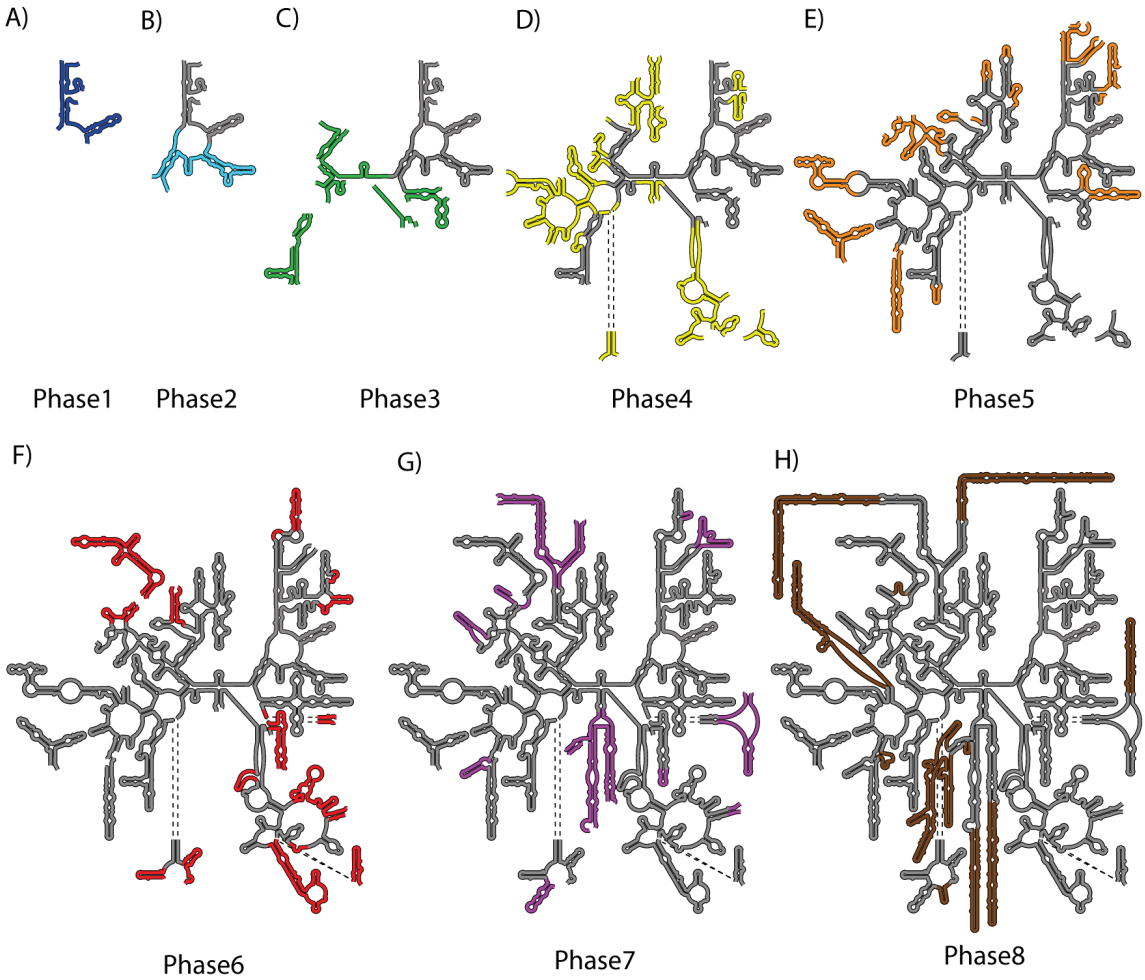

Figure 6.7. rRNA evolution mapped onto the LSU rRNA secondary structure. Accretion of ancestral and eukaryotic expansion segments is distributed into eight phases, associated with ribosomal functions: Phase 1) Rudimentary Binding and Catalysis, dark blue; Phase 2) Maturation of the PTC and Exit Pore, light blue; Phase 3) Early Tunnel Extension, green; Phase 4) Acquisition of the SSU Interface, yellow; Phase 5) Acquisition of Translocation Function, orange; Phase 6) Late Tunnel Extension, red; Phase 7) Encasing the Common Core (simple eukaryotes), purple; Phase 8) Surface elaboration (complex eukaryotes), brown.

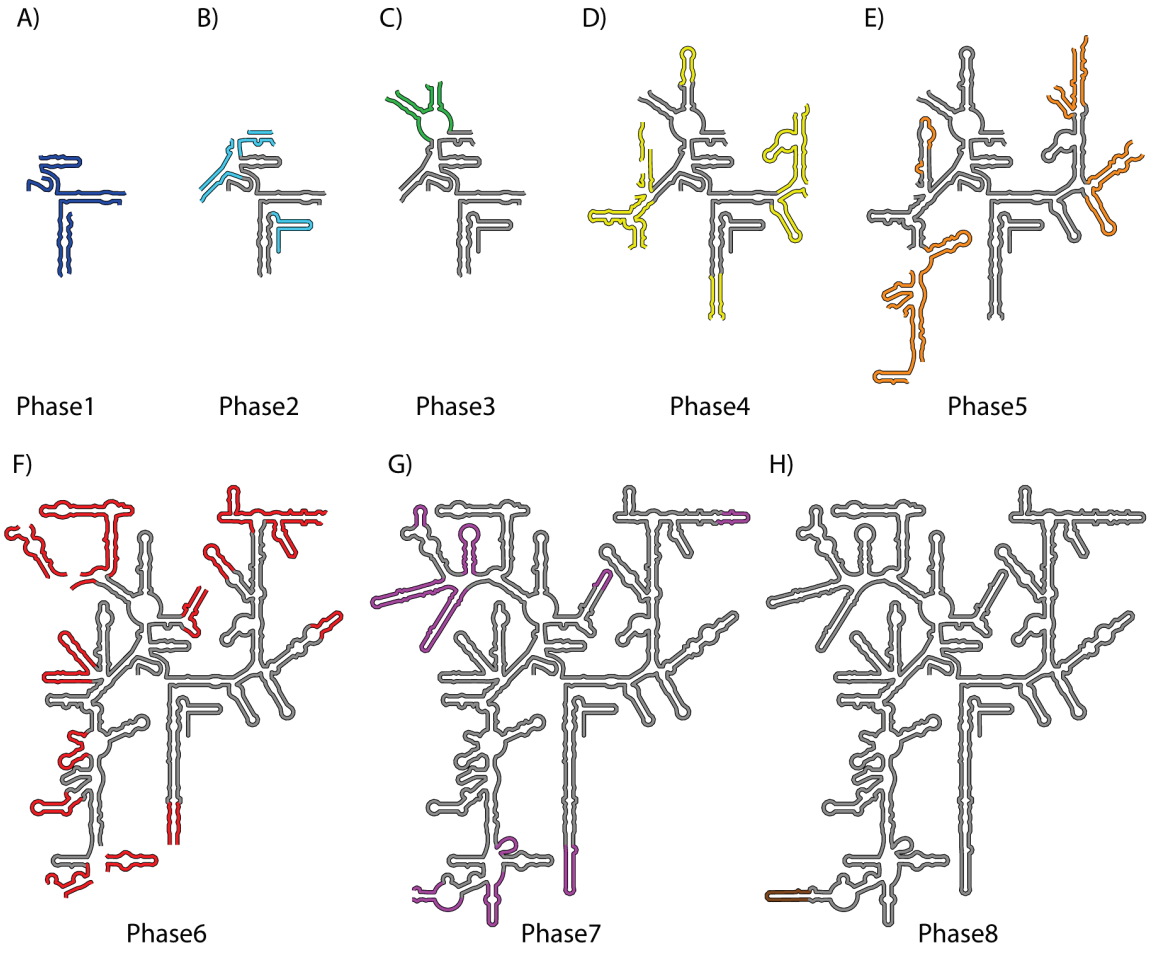

Figure 6.8. rRNA evolution mapped onto the SSU rRNA secondary structure. Accretion of ancestral and eukaryotic expansion segments is distributed into eight phases, associated with ribosomal functions: Phase 1) Origin of the decoding center, dark blue; Phase 2) Origin of the central pseudoknot and mRNA, light blue; Phase 3) Binding to the LSU, green; Phase 4) Stabilization of the LSU/SSU tRNA/mRNA complex, yellow; Phase 5) Origin of coding and translocation?, orange; Phase 6) Ribosomal tune up and surface decoration, red; Phase 7) Encasing the Common Core (simple eukaryotes), purple; Phase 8) Surface elaboration (complex eukaryotes), brown.

Phase 4) Acquisition of the SSU Interface. AES 11-28 are included. AES 11, 12, 15 form the LSU interface for association with the SSU. The other segments enhance the stability and efficiency of the LSU by embracing the PTC and further extending the exit tunnel.

Phase 5) Acquisition of Translocation Function. Inclusion of AES 29-39 adds essential components of the modern energy-driven translational machinery: the L7/L12 stalk and central protuberance,[142,143] and binding site (sarcin-ricin loop) for EF-G and EF-TU.[142,143] The tunnel is further extended.

Phase 6) Late Tunnel Extension. Further expansion of the LSU by inclusion of AES 40-59 results in the maturation of common core of the LSU. In the final phase of prokaryotic ribosomal evolution, the exit tunnel is extended. A majority of elements added here are located at the ribosomal surface and interact with ribosomal proteins

Phase 7) Encasing the Common Core (simple eukaryotes). Eukaryotic expansion segments are acquired and previous AES's are elongated. This eukaryotic-specific rRNA combines with eukaryotic-specific proteins[99] to form a shell around the common core.

Phase 8) Surface Elaboration (complex eukaryotes). Metazoan ribosomes are decorated with 'tentacle-like' rRNA elements that extend well beyond the subunit surfaces [57]. These tentacles (**Figure 5.17C**), are fundamentally different in structure and function than common core rRNA. Metazoan expansions appear to enable elaborate control, delivery and complexity, and are thought, for example to enable communication between the mRNA exit in the SSU

and the exit tunnel terminus in the LSU, and to facilitate interactions with eukaryotic-specific factors involved in membrane localization.

Expansion of the model to include the origin of the SSU decoding center and the emergence of ribosomal proteins is a work in progress. The driving force for the origin of the SSU before its use in the ribosome is unknown. The proto-SSU likely had other uses in the prebiotic RNA-based World. The ancestors of the rProteins would be randomly synthesized peptide / polyester like molecules, fulfilling similar roles of rRNA stabilization. The rProteins have been coevolving along with the rRNA and should not be ignored in a complete model.

6.5 Integrated model of Ribosomal Evolution

We believe that the LSU and SSU rRNAs have separate origins from the same prebiotic world. Relatively little is known about the prebiotic world. While, we may never know exactly how modern day biochemistry evolved from a prebiotic chemistry world, progress is being made on providing a plausible pathway. [144-151] If a resurrected ancestral ribosome works with prebiotic chemistry,[152-154] the model is one step closer to plausible.

The relative timing of the first three phases of the LSU and SSU are unknown, but it is believed that they both came together at their respective phase 4. Before phase 4, the LSU and SSU might have interacted through a 3rd party, such as the ancestor to tRNA, but they had no predicted direct interactions. During phase 4, inter-subunit bridges formed, which either enhanced interactions through a 3rd party molecule, or enabled them for the first time. Either way, at phase 4, the LSU and SSU would have certainly coevolved.

We illustrate phases 1-4 of ribosomal RNA evolution using the secondary (**Figure 6.9**) and 3D (**Figure 6.10**) structures of *Homo sapiens*. Human structures are used to be consistent with later 8 phase figures. For the first 4 phases, the structure of the resulting rRNA would be very similar regardless of which species is used. In the secondary structures, it looks like the RNA is composed of several strands and the ends are far apart. This is an illusion caused by the shape of the full secondary structures. In 3D, the separate RNA strands could be ligated as is or with the addition of just one nucleotide. By Phase 3, the functional and structural centers of the LSU and SSU have formed. In Phase 4, inter-subunit bridges are formed and the complex is stabilized. Starting with Phase 4, the timelines for evolution are the same, and the molecules coevolve.

We illustrate phases 4-6 of ribosomal RNA evolution using the secondary (**Figure 6.11**) and 3D (**Figure 6.12**) structures of *Homo sapiens*. Phase 4 is the first true ancestral ribosome. During phase 5, the translocation machinery is developed. The L7/L12 stalk is formed, which helps bring A-site tRNA into the ribosome. The L7/L12 stalk forms the GTPase center along with ribosomal proteins. The elongation factors EF-Tu and EF-G bind here.[155] The 5S appears, which forms the central protuberance and helps coordinate ribosome functions.[156] The L1 stalk forms, which helps more tRNA from the P-site to the E-site and general translocation mechanisms.[157] Phase 6 completes the common core and modern prokaryotic ribosomes.

We illustrate phases 6-8 of ribosomal RNA evolution using the secondary (**Figure 6.13**) and 3D (**Figure 6.14**) structures of *Homo sapiens*. Phase 7 marks the transition to eukaryotic ribosomes. Phase 8 is the development of long eukaryotic expansion segments associated with higher eukaryotes, especially birds and mammals.

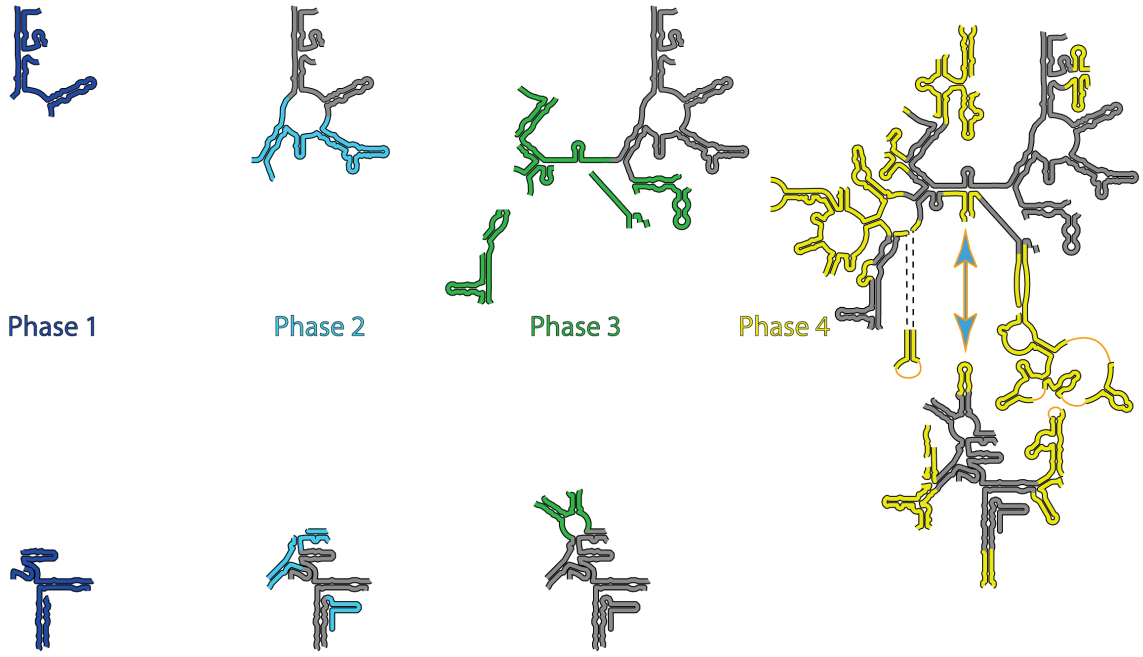

Figure 6.9. Integrated model of ribosomal evolution, phases 1-4. During phases 1-3, the interaction between the LSU and SSU, if any, are unknown. The timeline of the phases may be out of sync. However, at phase 4, the timelines for the LSU and SSU converge, and inter-molecular bridges (not shown) form. *H. sapiens* rRNA is used for illustration.

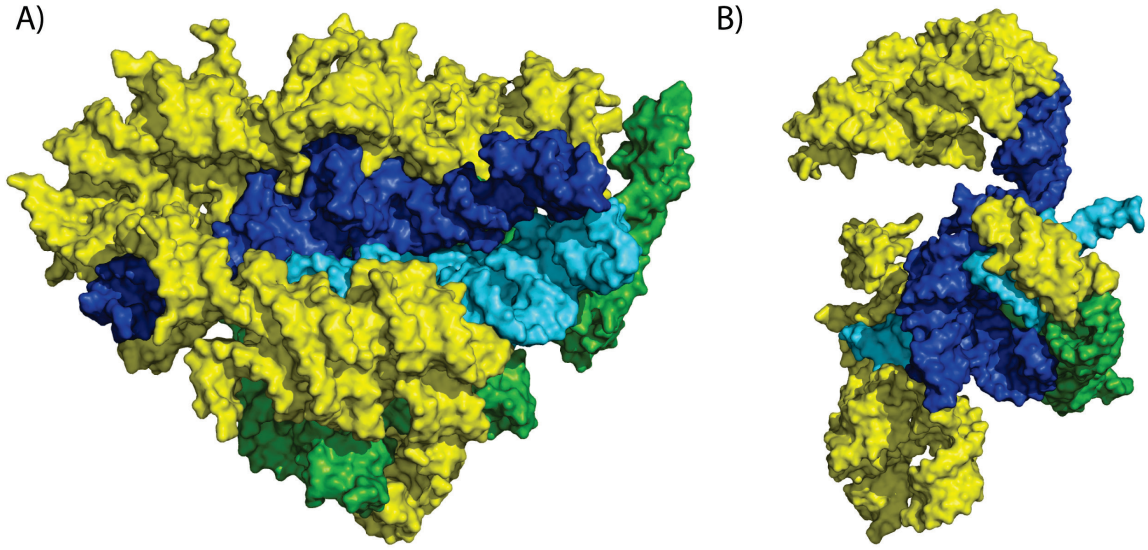

Figure 6.10. 3D images of the A) LSU, and B) SSU, at phase 4. *H. sapiens* rRNA is used for illustration. Phase 1 is dark blue, Phase 2 is light blue, Phase 3 is green, and Phase 4 is yellow. It is at phase 4 that inter-subunit bridges appear. It is likely that the tRNA ancestor has formed its characteristic shape by Phase 4, if not earlier.

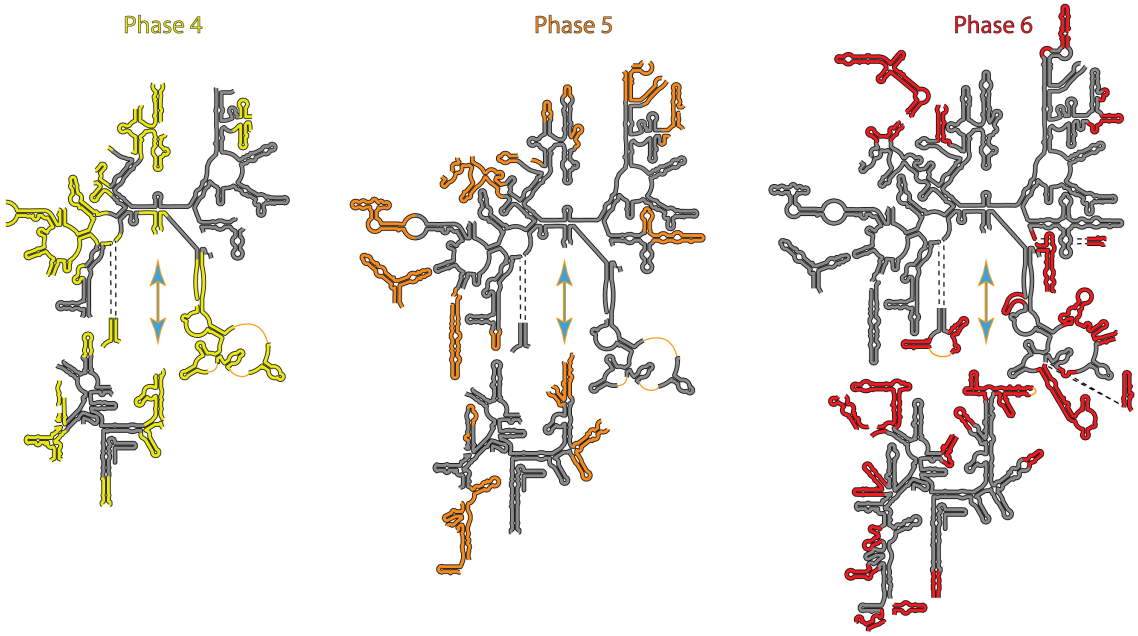

Figure 6.11. Integrated model of ribosomal evolution, phases 4-6, mapped on the secondary structure of *H. sapiens* rRNA. At phase 4, the timelines for the LSU and SSU converge, and inter-molecular bridges (not shown) form. Phase 5 marks the advent of translocation machinery. The L7/L12 stalk, L1 stalk, and central protuberance (5S) form. Phase 6 marks the completion of the common core.

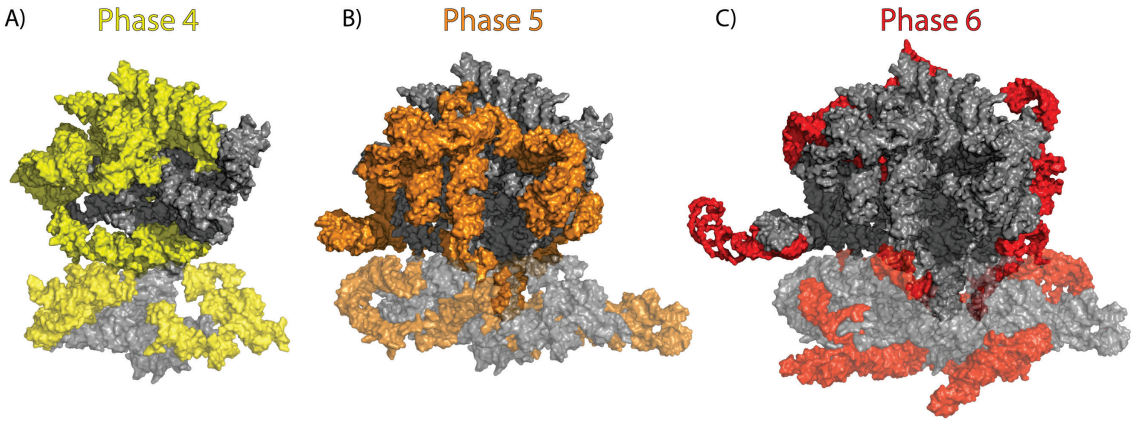

Figure 6.12. Integrated model of ribosomal evolution, phases 4-6, mapped on the 3D structure of *H. sapiens* rRNA. At phase 4, the timelines for the LSU and SSU converge, and inter-molecular bridges (not shown) form. Phase 5 marks the advent of translocation machinery. The L7/L12 stalk, L1 stalk, and central protuberance (5S) form. Phase 6 marks the completion of the common core. The SSU is slightly transparent and uses slightly different shades of coloring, than the LSU.

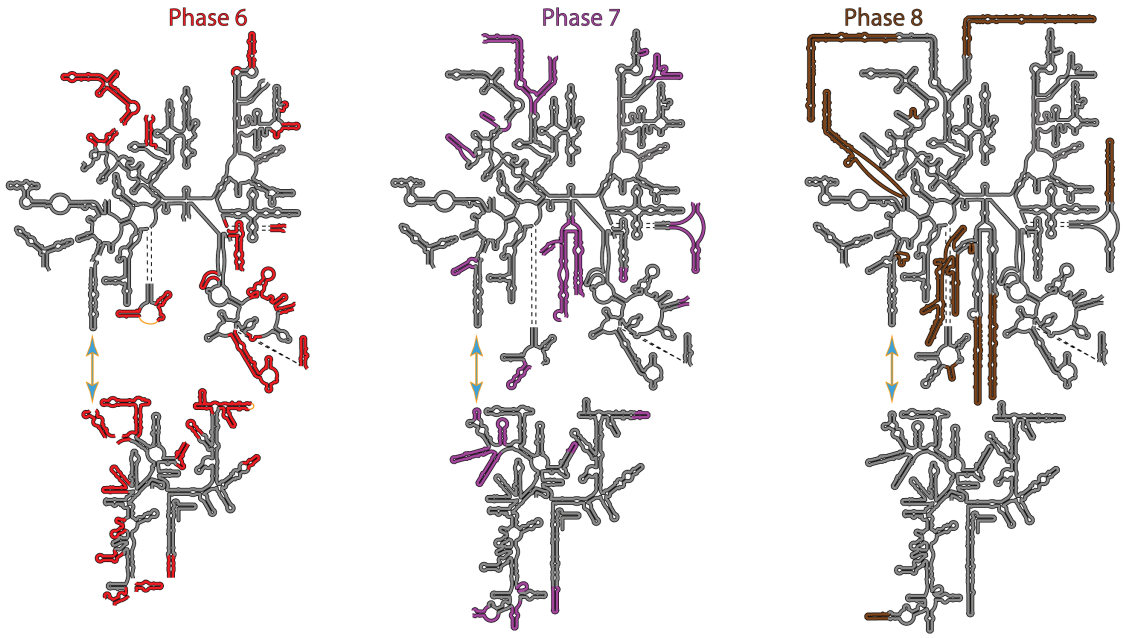

Figure 6.13. Integrated model of ribosomal evolution, phases 6-8, mapped on the secondary structure of *H. sapiens* rRNA. Phase 6 marks the completion of the common core. Phase 7 contains most eukaryotic expansion segments. In Phase 8, the eukaryotic expansion segments get significantly longer. Phase 8 corresponds with the emergence of higher eukaryotes such as birds and mammals.

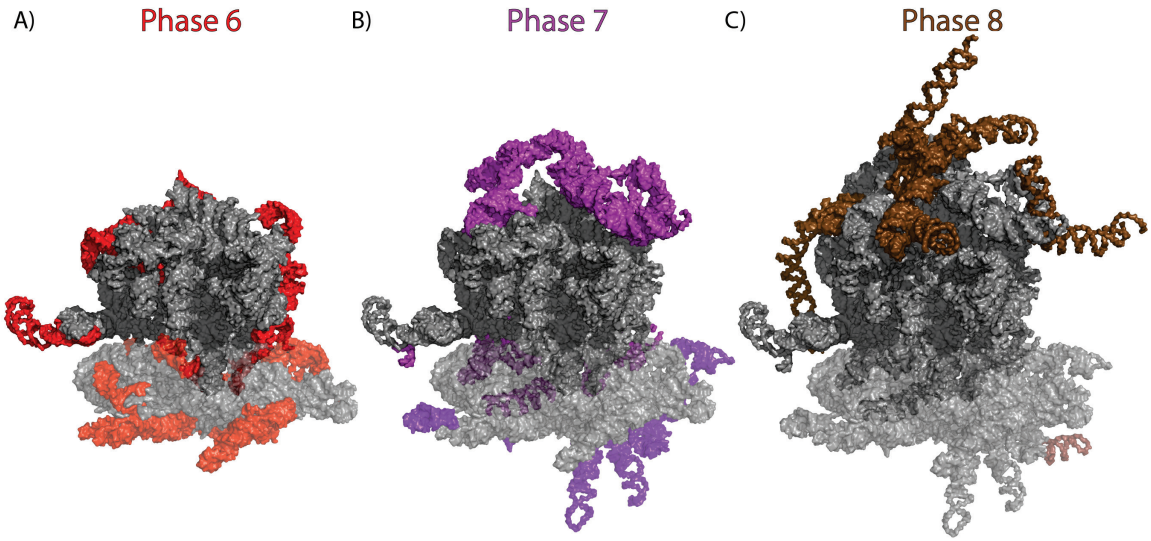

Figure 6.14. Integrated model of ribosomal evolution, phases 6-8, mapped on the 3D structure of *H. sapiens* rRNA. Phase 6 marks the completion of the common core. Phase 7 contains most eukaryotic expansion segments. In Phase 8, the eukaryotic expansion segments get significantly longer. Phase 8 corresponds with the emergence of higher eukaryotes such as birds and mammals.

6.6 Discussion

We have introduced the concept of an *'insertion fingerprint'*. Insertion fingerprints are a pattern of rRNA structure where it looks as if a trunk helix has been inserted into a branch helix, with as little perturbation of the trunk as possible. We have demonstrated that insertion fingerprints are a widely distributed phenomenon. Insertion fingerprints appear in every part of the ribosome and date back to its earliest origins.

Insertion fingerprints provided a convenient way to dissect the ribosome into small manageable pieces, called ancestral expansion segments (AES's). Most AES's should fold independently, at least at the secondary structure level. Experimentally, it has been shown that the entire Domain III can fold independently.[158] It is likely that AES's in the inner core of the ribosome do not require the presence of AES's on the outside.

AES's immediately suggest a way to synthesize subsets of the ribosome for *in vivo* or *in vitro* studies. Most AES's can be removed without changing the remaining RNA. It would not be necessary to add any unnatural RNA sequence, whether single-stranded or helical.

AES's provide a better way to think about the structure of the ribosome. Traditionally defined helices only refer to the actual helical regions. The remaining "single-stranded" regions do not have a formal name. Some of these regions are functionally important, highly conserved, and/or not actually single-stranded.

We presented a model of ribosomal evolution based on AES's. The model obeys the Darwin continuity principle from the first addition. There is a single origin of the LSU and a single origin of the SSU. While it's possible that larger piece had evolved and fused together, this is not necessary for proper functioning of the ribosome. Our model is

the first to explain the origin of the PTC and begins with a small piece easily achievable under random RNA synthesis on the pre-biotic Earth.

We present the first model for SSU evolution and place it approximately on the same timescale as the LSU. The model can be resurrected in the laboratory. If an appropriate activity assay can be developed, a lot of hypotheses could be tested. Information can be gained not only about the rRNA itself, but everything associated with it. Different polymers and monomers could be tried. The origin of mRNA and tRNA could be studied. Ancestral ignition and elongation factors could be tested. Eventually, it may be possible to artificially evolve not only the entire translation system, but a simple organism.

CHAPTER 7

DISCUSSION

7.1 Discuss usefulness of RiboZones and our Philosophy

RiboZones data, analysis tools, and visualization tools will continue to develop, and become a valuable resource to the ribosomal community. The community should add sequences, structures, and features. There is a tremendous amount of information already known spread throughout the literature. RiboZones has a goal of inspiring the community to organize and catalog the known data.

RiboVision has been released to the public for well over a year. During that time, we have tracked hits from 50 countries and 394 cities. While many hits are cursory, some are active repeat users. The next release of RiboVision will contain more detailed tracking statistics.

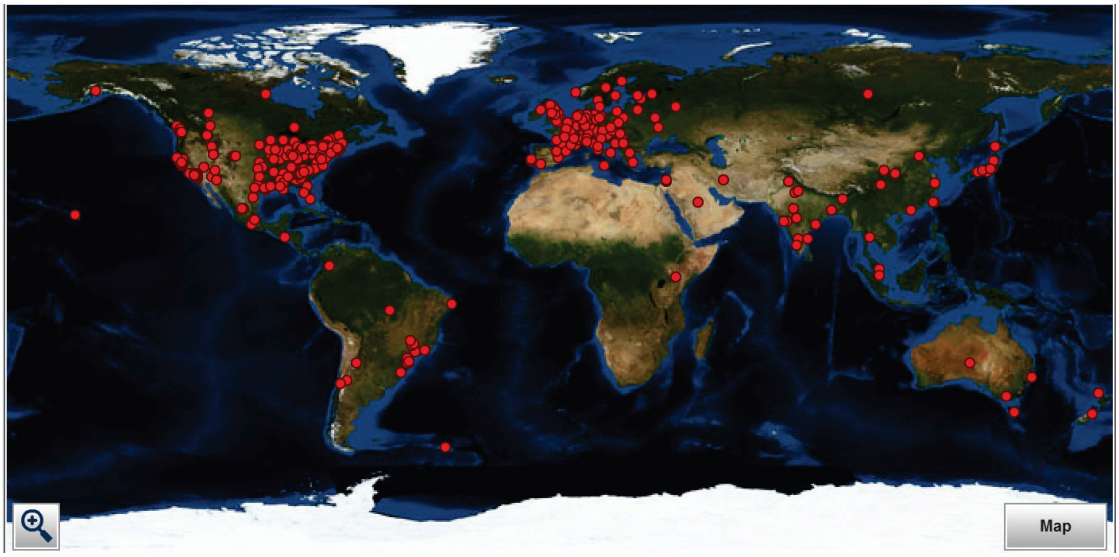

Figure 7.1. Map of hits for the RiboVision web portal. Hits are from 50 countries and 394 cities.

RiboVision has proven incredibly useful to our lab and collaborators. Every feature enables new use cases. As data and features increase, the ability to do "exploratory research" increases. Some simple questions can be answered in minutes, and more complicated ones can often be answered in a few hours or days. The ease of use allows preliminary data to be visualized just as well as final data.

The ResidueTip feature is unique. Currently, the ResidueTip window shows nucleotide name, consensus sequence, Shannon entropy, and nucleotide frequency distributions. ResidueTip also shows "Selected Data", which could be from any layer. This feature allows visualization of multiple variables simultaneously and will become more useful as the feature expands.

RiboVision provides protection against future changes to secondary structure diagrams. Saving figures in RiboVision CSV format allows quick regeneration onto a new secondary structure in the future. We have changed the secondary structure diagrams substantially for the LSU and moderately for the SSU already. We never would have changed the secondary structure diagrams without RiboVision. As this dissertation shows, the secondary structures are not as accurate as they could be. The secondary structures may be changed again in the future, but storing data in RiboVision format will allow quick regeneration of new versions of the figures.

RiboVision is a young program and its true power will only be realized in future versions. Short-term improvements include an interactive sequence viewer and more data sharing from the 3D side back to the 2D side. Medium term improvements would include an optional user account system, with all the features that implies, easier ability for users to submit data, and structure comparison features. At first, intra species comparisons will

be developed, to compare data from the same species, but perhaps in different states. Later, inter-species features will be developed. It is hoped that the community will contribute data, ideas, and even code.

We believe that RiboVision reduces the gap between experimental scientists and computational scientists. There is a heavy focus on user-friendliness and a shallow learning curve. Computational scientists can better focus on developing novel algorithms. Experimental scientists can become more closely acquainted with their data, doing more basic level bioinformatics analysis, and developing their new data into higher significance.

7.2 Discuss Alignment

We have developed the most complete and accurate MSA of SSU and LSU rRNA sequences available. The sample size, 133 is relatively small, but adding more sequences is relatively simple. MAFFT has been shown to work well using the RiboZones MSA as a template. Getting the alignment to such a high level of completeness was a lot of work. Some species were very well annotated, and had complete rRNA sequences available. Some sequences took several days of effort each. Their sequences could only be reconstructed after searching multiple databases with multiple queries and tools, and piecing together data from several sources. This is not ideal, but it is better than incomplete sequences.

Our resulting MSA has several features that are not common to other researchers' alignments. We believe that these features enable new kinds of analysis with less bias and more statistical certainty.

We provide a matching set of alignments for the LSU and SSU. Often alignments are only made for one subunit. When alignments are made for both subunits, they typically have different species lists. A common strategy for making an alignment is taking all available data from a source through a particular query method, and then applying a particular filtering method. This will certainly result in a different number and distribution of LSU sequences vs SSU sequences. There is an order of magnitude more SSU sequences available than LSU sequences. Sequencing the SSU of every known species is routine analysis, part of defining the phylogeny and the name of a species. Therefore, the distribution of SSU sequences should be a relatively fair representation of the entire biodiversity of Earth. Much fewer LSU sequences are known. The species chosen are heavily biased. Species chosen for LSU sequencing are more likely to be model organisms, species of economic interest, or pathogens. By providing a matching set, we remove the variables of different species and different number of species. This makes broad comparisons between the subunits more valid. It also ensures that inter-subunit comparisons can be made. Predictions about correlations between sequence or structure changes between subunits can be tested.

Our sequence list contains only fully sequenced organisms. The biggest benefit of this is that a matching MSA can be made for any protein with homologs in all three domains of life. Fair analysis can be performed studying the relationship between any rRNA and protein, for example, the 16S/18S rRNA and EF-Tu, or the LSU and trigger factor. Additionally, analysis can be performed using additional genome level variables. A fully sequenced organism has additional information available, such as c-value, number of chromosomes, presence of other genes, overall GC content, etc.

Our MSA has been adjusted for 3D structure in many places, making our alignment better suited for homology modeling. Our MSA can be used to create a "correspondence table." A correspondence table is a one-to-one mapping of structure aware sequence between two species. The MSA needs more work before it can produce a perfect correspondence table, but it is the most accurate so far. Uses of correspondence tables include making a new secondary structure, and converting data between different species numbering systems. A perfect correspondence table would allow automating these procedures, which would be highly valuable features for RiboVision.

Our alignment is complete. A complete MSA allows for even visualization across entire rRNA molecules. Statistics for Domains I and VI can fairly be compared to statistics from Domain IV and V. Incomplete sequences reduce the signal and increase the noise, obscuring what could be significant observations, and potentially causing false conclusions. A complete alignment is a requirement for making a nucleotide level definition of the universal common core. If, for example, 20% of sequences were incomplete, then it would be impossible to define an accurate common core to the 95% level. Large sections of the ribosome would be marked as not common core. As a side effect, there would probably be some helices that had one side in the common core and one side not in the common core. The alternative, of reducing the common core cutoff to be below 80% also produces an inaccurate model. Nucleotides that were genuinely only present in 80% of species would be marked the same as nucleotides that only appear to be present in 80% of species due to incomplete data. The differences between the two types would be technically indistinguishable. Additionally, if genuinely only present less than

80%, does that deserve to be called common? The usefulness of the common core would be decreased dramatically.

7.3 Discuss Common Core

We have made the first statistically validated nucleotide-level definition of the common core. Previously, the common core was only discussed as an abstract concept and cartoon level visualization. We have defined multiple versions of the common core: universal common core, prokaryotic common core, bacterial common core, archaeal common core, and eukaryotic common core. The appropriate common core models can be chosen to match the type of analysis desired.

The prokaryotic common core is a first approximation of the ribosome of LUCA. Models of LUCA ribosomes can be resurrected for use in *in vitro* or *in vivo* experiments. The prokaryotic common core defines what part of the rRNA should definitely be included in LUCA. The remaining helices could be optionally included as part of experiment design. Comparisons could be made between three versions of the LUCA model, for a particular helix, a version with (i) a bacterial helix, (ii) an archaeal helix, (iii) minimum length, possibly 0, helix. Experiments would lead to a definition and characterization of LUCA.

Prokaryotic common core analysis, LUCA models, etc., could help root the tree of life, a big open question. It is unknown where LUCA falls on the spectrum from bacteria to archaea. It is necessary to root the tree of life based on external data.[108,159,160] Traditionally, the root is placed near bacteria, but some evidence shows that the root is closer to archaea.[105] A very detailed understanding of ribosomal structure and function may help root the tree of life with higher confidence.

Domain specific common core analysis could help the differences between the domains be characterized and possibly explained. The three domains of life have many base pairs that are preserved within a domain, but different between them. For example, in some places, a base pair might always be a GC in bacteria, but a CG in archaea. In other places, it could be a CG or GC in bacteria, but always an AU or UA in archaea. Detailed analysis and visualization could identify the pattern for each base pair. This could eventually lead to a good model for the consensus sequence of the common core and a root for the tree of life.

The origin of eukaryotes could be studied. What is the sequence and structure of LECA? What caused eukaryotes to separate from archaea? Why did eukaryotes evolve from archaea and not bacteria? Can bacterial ribosomes grow large, and if not, what's preventing that?

The ribosomal proteins could be studied. Did the proteins diverge before or after the rRNA? Are these events independent, correlated, or causative? Is it possible to form hybrid ribosomes that are part archaea and part eukarya?

Analyzing common models would help drug discovery. For example, finding differences between the bacterial common core and the eukaryotic common core could help identify potential anti-biotic binding sites. Similarly, anti-fungals, insecticides, anti-malarials, etc could be screened for. Trypanosomes have very different ribosomes, so maybe a drug could be discovered to interfere with trypanosomal ribosome function.

7.4 Discuss Model

We have proposed a tentative model of evolution of the entire ribosome from its very first simple element. The ribosome continually grew through accretion of RNA

elements called ancestral expansion segments (AES's). At first it grew extremely quickly, developing from the first element up through the common core, in a time span of up to 500 million years, possibly much faster. The ribosome's growth reached a plateau for over a billion years. Then, around 2 billion years ago, something triggered the development of eukaryotes. As eukaryotes developed higher levels of complexity, so did their ribosomes.

In our model, the LSU has evolved in distinct phases. This process started with the formation of the P-site, possibly in an RNA world, and continues today in eukaryotes. A unifying theme of LSU evolution is the continuous extension, stabilization and elaboration of exit tunnel structure and function. The exit tunnel is formed, extended, stabilized, and elaborated continuously in nearly all phases of ribosomal evolution.

The model of LSU origins and evolution described here is more fine-grained than previous models but is in essential agreement with them, despite different assumptions and types of input data. Harvey and co-workers compared secondary structures and sequences across multiple species, identifying the RNA components of the "minimal ribosome." [134] Fox analyzed density of molecular interactions and interconnectivities.[109] Bokov and Steinberg developed a powerful model by analyzing A-minor interactions.[138] Williams and coworkers treated the LSU as a growing onion.[58] Where they overlap, our stepwise model here corresponds well with each of these previous models, although it provides a more rigorous definition of the ancestral expansion segments and addresses the origin of the PTC. The cumulative effect of the first four initial expansions (**Figure 6.6**), gives a structure that is strikingly similar to an ancestral PTC proposed independently by Yonath and coworkers.[161,162] Those investigators suggested rRNA

components of the PTC as an ancient catalytic heart of the common core. Some of the AES's proposed here correspond to rRNA 'elements' that were used to construct the ribosome in the Bokov-Steinberg model.[138]

Our model is the first to define non-arbitrary boundaries between elements. Removing one AES does not change the structure of the remaining rRNA. Unnatural tetraloops or other sequence does not need to be added. Each AES should be roughly structurally independent. As such, our model is convenient to test *in vitro*.

Our model proposes increased functionality upon addition of each AES, or group of AESs. Not only did the structure of the ribosome grow in a logical manner, so did its functionality. If functionality grew in an unexplainable order, it would not be the most likely model. First, the catalytic center evolved. Improvements to the catalytic core include development of an exit tunnel and possibly binding to something like the precursor to tRNA. The ancestral SSU may stabilize the LSU-tRNA complex, providing the evolutionary pressure to develop better interaction between the LSU and the SSU to further stabilize the complex. The tRNA molecule matures during this time. With the tRNA molecule shape mature, the translocation machinery can develop. The result is a dramatically increased rate of reaction, allowing all kinds of proteins to evolve. Only at the end, does accessory functionality, such as association with chaperones and transporters develop.

Each step of the model can be tested in the laboratory. The model can be revised as necessary. In the end, we gain not just a model of ribosomal evolution, but have come to thoroughly understand the function of the ribosome.

The biggest event in the history of the ribosome is the transition from non-coded to coded peptide synthesis. That event is the true origin of life and the true beginning of biology. Like every other complex event in biology, it would have happened through a series of much smaller steps, each with their own selective advantage. It would be good to know the most likely pathway of how coding developed.

Once the ideal conditions are known, it may be possible to accelerate evolution and evolve a whole ribosome in a reasonable time frame, say 10 years. If this is possible, alternative ribosomes may also develop. It would be highly informative to learn what alternatives to Earth ribosomes are possible. If life had evolved on a different planet, how similar would the life there be to Earths?

7.5 Conclusions

RiboZones is a proven useful suite of data and tools. They are designed for a broad variety of research applications. Here, we presented a small, but powerful, fraction of applications possible. A high-quality MSA was built using special conditions to give it more research power. A detailed model of the universal common core and variations on it was developed. The ribosome was dissected into useful structural and evolutionary units, AES's. AES's provided a new way to organize the structure of the ribosome, and suggested a useful evolutionary model.

RiboZones will continue to be developed. There are dozens more crystal structures to process, however mostly of just a few species. There are differences between them though. They have been crystallized with different proteins and small molecules, trapping the structures in different states. Comparative analysis tools will be

developed and generalized to show these differences. Progress is underway in adding mitochondrial structures to RiboZones.

In the near future, emphasis will be placed on building a community around RiboZones. Features which facilitate user sharing of data and code will be developed. Support forums, discussion forums, and wiki's will be started. The usefulness of RiboZones goes up exponentially with the number of users.

Work is underway on better understanding the role of rProteins and integrating them into the evolutionary models. The RiboZones model will be the first to integrate a detailed history of the rProteins into the model.

The secondary structure for both the LSU and SSU should once again be majorly revised. The current structures have their place and will stay supported for the foreseeable future. When just trying to show new data about a particular part of the ribosome, these structures work fine. However, highly accurate alternative versions need to be created for use in studies that consider the structure and evolution of whole ribosomes. RiboVision makes working with secondary structures so easy, alternative versions of secondary structures are no longer a problem.

In sum, we have tamed the ruthless ribosome. We have simplified the process of starting an initial dataset. We have made visualization of data fast and efficient, so that it is no longer a bottleneck. We have made the most accurate, complete, three domain of life sequence alignment. We have statistically defined the common core of the ribosome for all life. We have partitioned the ribosome into useful units, AES's. Everything here lead to the first complete, continuous evolution, experimentally testable, model for rRNA evolution. Finally, theories about the origin of life are more testable.

REFERENCES

1. Wolf, Y. I.; Koonin, E. V. *Biol Direct* **2007**, *2*, 14.

2. Bieling, P.; Beringer, M.; Adio, S.; Rodnina, M. V. *Nature structural & molecular biology* **2006**, *13*, 423.

3. Schmeing, T. M.; Huang, K. S.; Strobel, S. A.; Steitz, T. A. *Nature* **2005**, *438*, 520.

4. Weiss, R.; Cherry, J. *The RNA World* **1993**, 71

5. Pestka, S. *J Biol Chem* **1969**, *244*, 1533

6. Lee, J. C.; Gutell, R. R. *PLoS ONE* **2012**, *7*, e38203.

7. Ramakrishnan, V. *Cell* **2002**, *108*, 557.

8. Gualerzi, C. O.; Pon, C. L. *Biochemistry* **1990**, *29*, 5881.

9. Gribskov, M. *Gene* **1992**, *119*, 107.

10. Spedding, G.; Gluick, T. C.; Draper, D. E. *J. Mol. Biol.* **1993**, *229*, 609.

11. Poole, E. S.; Brimacombe, R. *RNA* **1997**, *3*, 974.

12. Petry, S.; Weixlbaumer, A.; Ramakrishnan, V. *Current Opinion in Structural Biology* **2008**, *18*, 70.

13. Korostelev, A. A. *RNA* **2011**, *17*, 1409.

14. Nissen, P.; Hansen, J.; Ban, N.; Moore, P. B.; Steitz, T. A. *Science* **2000**, *289*, 920.

15. Rohl, R.; Nierhaus, K. H. *Proc Natl Acad Sci U S A.* **1982**, *79*, 729.

16. Schulze, H.; Nierhaus, K. H. *EMBO J* **1982**, *1*, 609.

17. Nitta, I.; Kamada, Y.; Noda, H.; Ueda, T.; Watanabe, K. *Science* **1998**, *281*, 666.

18. Cooperman, B. S.; Wooten, T.; Romero, D. P.; Traut, R. R. *Biochem Cell Biol.* **1995**, *73*, 1087.

19. Uhlein, M.; Weglöhner, W.; Urlaub, H.; Wittmann-Liebold, B. *Biochem J.* **1998**, *331*, 423.

20. Diedrich, G.; Spahn, C. M.; Stelzl, U.; Schafer, M. A.; Wooten, T.; Bochkariov, D. E.; Cooperman, B. S.; Traut, R. R.; Nierhaus, K. H. *EMBO Journal* **2000**, *19*, 5241.

21. Willumeit, R.; Forthmann, S.; Beckmann, J.; Diedrich, G.; Ratering, R.; Stuhrmann, H. B.; Nierhaus, K. H. *J Mol Biol* **2001**, *305*, 167.

22. Chodavarapu, S.; Felczak, M. M.; Kaguni, J. M. *Nucleic Acids Res* **2011**, *39*, 4180.

23. Petrov, A. S.; Bernier, C. R.; Hsiao, C.; Okafor, C. D.; Tannenbaum, E.; Stern, J.; Gaucher, E.; Schneider, D.; Hud, N. V.; Harvey, S. C.; Williams, L. D. *J. Phys. Chem. B* **2012**, *116*, 8113.

24. Warner, J. R.; McIntosh, K. B. *Molecular Cell*, *34*, 3.

25. Vilardell, J.; Warner, J. R. *Mol Cell Biol* **1997**, *17*, 1959.

26. Carbon, P.; Ehresmann, C.; Ehresmann, B.; Ebel, J. P. *Eur J Biochem.* **1979**, *100*, 399.

27. Klindworth, A.; Pruesse, E.; Schweer, T.; Peplies, J.; Quast, C.; Horn, M.; Glöckner, F. O. *Nucleic Acids Research* **2013**, *41*, e1.

28. Gillespie, J. J.; Johnston, J. S.; Cannone, J. J.; Gutell, R. R. *Insect Mol. Biol.* **2006**, *15*, 657.

29. Galperin, M. Y.; Koonin, E. V. *Trends Biotechnol* **2010**, *28*, 398.

30. Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. *Nucleic Acids Res* **2013**, *41*, D36.

31. Leinonen, R.; Akhtar, R.; Birney, E.; Bower, L.; Cerdeno-Tarraga, A.; Cheng, Y.; Cleland, I.; Faruque, N.; Goodgame, N.; Gibson, R.; Hoad, G.; Jang, M.; Pakseresht, N.; Plaister, S.; Radhakrishnan, R.; Reddy, K.; Sobhany, S.; Ten Hoopen, P.; Vaughan, R.; Zalunin, V.; Cochrane, G. *Nucleic Acids Res* **2011**, *39*, D28.

32. Logan-Klumpler, F. J.; De Silva, N.; Boehme, U.; Rogers, M. B.; Velarde, G.; McQuillan, J. A.; Carver, T.; Aslett, M.; Olsen, C.; Subramanian, S.; Phan, I.; Farris, C.; Mitra, S.; Ramasamy, G.; Wang, H.; Tivey, A.; Jackson, A.; Houston, R.; Parkhill, J.; Holden, M.; Harb, O. S.; Brunk, B. P.; Myler, P. J.; Roos, D.; Carrington, M.; Smith, D. F.; Hertz-Fowler, C.; Berriman, M. *Nucleic Acids Res* **2012**, *40*, D98.

33. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res* **1997**, *25*, 3389.

34. Kent, W. J. *Genome Res* **2002**, *12*, 656.

35. Cole, J. R.; Chai, B.; Marsh, T. L.; Farris, R. J.; Wang, Q.; Kulam, S. A.; Chandra, S.; McGarrell, D. M.; Schmidt, T. M.; Garrity, G. M.; Tiedje, J. M. *Nucleic Acids Res.* **2003**, *31*, 442.

36. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F. O. *Nucleic Acids Research* **2013**, *41*, D590.

37. DeSantis, T. Z.; Hugenholtz, P.; Larsen, N.; Rojas, M.; Brodie, E. L.; Keller, K.; Huber, T.; Dalevi, D.; Hu, P.; Andersen, G. L. *Appl. Environ. Microbiol.* **2006**, *72*, 5069.

38. Cannone, J. J.; Subramanian, S.; Schnare, M. N.; Collett, J. R.; D'Souza, L. M.; Du, Y.; Feng, B.; Lin, N.; Madabusi, L. V.; Muller, K. M.; Pande, N.; Shang, Z.; Yu, N.; Gutell, R. R. *BMC Bioinformatics* **2002**, *3*, 2.

39. Branlant, C.; Krol, A.; Machatt, M. A.; Pouyet, J.; Ebel, J. P.; Edwards, K.; Kossel, H. *Nucleic Acids Research* **1981**, *9*, 4303.

40. Zuker, M. *Curr Opin Struct Biol* **2000**, *10*, 303.

41. Hofacker, I. L.; Fekete, M.; Stadler, P. F. *J Mol Biol* **2002**, *319*, 1059.

42. Hofacker, I. L. *Nucleic Acids Research* **2003**, *31*, 3429.

43. Petrov, A. S.; Bernier, C. R.; Hershkovits, E.; Xue, Y.; Waterbury, C. C.; Hsiao, C.; Stepanov, V. G.; Gaucher, E. A.; Grover, M. A.; Harvey, S. C.; Hud, N. V.; Wartell, R. M.; Fox, G. E.; Williams, L. D. *Nucleic Acids Res* **2013**, *41*, 7522.

44. Petrov, A. S.; Bernier, C. R.; Gulen, B.; Waterbury, C. C.; Hershkovits, E.; Hsiao, C.; Harvey, S. C.; Hud, N. V.; Fox, G. E.; Wartell, R. M.; Williams, L. D. *PLoS ONE* **2014**, *9*, e88222.

45. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *Journal of Molecular Biology* **1977**, *112*, 535.

46. Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S.-H.; Srinivasan, A. R.; Schneider, B. *Biophys. J.* **1992**, *63*, 751.

47. Jarasch, A.; Dziuk, P.; Becker, T.; Armache, J. P.; Hauser, A.; Wilson, D. N.; Beckmann, R. *Nucleic Acids Res* **2012**, *40*, D495.

48. Petrov, A. I., Bowling Green State University, 2012.

49. Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. *Bioinformatics* **2007**, *23*, 2947.

50. Li, H.; Homer, N. *Briefings in Bioinformatics* **2010**, *11*, 473.

51. Smith, C.; Heyne, S.; Richter, A. S.; Will, S.; Backofen, R. *Nucleic Acids Res* **2010**, *38*, W373.

52. Bernier, C. R.; Petrov, A. S.; Waterbury, C. C.; Jett, J.; Li, F.; Freil, L. E.; Xiong, X.; Wang, L.; Migliozzi, B. L. R.; Hershkovits, E.; Xue, Y.; Hsiao, C.; Bowman, J. C.; Harvey, S. C.; Grover, M. A.; Wartell, Z. J.; Williams, L. D. *Faraday Discussions* **2014**.

53. Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. *Science* **2000**, *289*, 905.

54. Selmer, M.; Dunham, C. M.; Murphy, F. V.; Weixlbaumer, A.; Petry, S.; Kelley, A. C.; Weir, J. R.; Ramakrishnan, V. *Science* **2006**, *313*, 1935.

55. Dunkle, J. A.; Wang, L. Y.; Feldman, M. B.; Pulk, A.; Chen, V. B.; Kapral, G. J.; Noeske, J.; Richardson, J. S.; Blanchard, S. C.; Cate, J. H. D. *Science* **2011**, *332*, 981.

56. Ben-Shem, A.; de Loubresse, N. G.; Melnikov, S.; Jenner, L.; Yusupova, G.; Yusupov, M. *Science* **2011**, *334*, 1524.

57. Anger, A. M.; Armache, J. P.; Berninghausen, O.; Habeck, M.; Subklewe, M.; Wilson, D. N.; Beckmann, R. *Nature* **2013**, *497*, 80.

58. Hsiao, C.; Mohan, S.; Kalahar, B. K.; Williams, L. D. *Mol Biol Evol* **2009**, *26*, 2415.

59. http://rna.ucsc.edu/rnacenter/xrna/xrna.html.

60. Darty, K.; Denise, A.; Ponty, Y. *Bioinformatics* **2009**, *25*, 1974.

61. Noller, H.; University of California, Santa Cruz; Vol. 2014.

62. Pruesse, E.; Peplies, J.; Glöckner, F. O. *Bioinformatics* **2012**, *28*, 1823.

63. Penn, O.; Privman, E.; Ashkenazy, H.; Landan, G.; Graur, D.; Pupko, T. *Nucleic Acids Research* **2010**, *38*, W23.

64. Hsiao, C. L.; Mohan, S.; Kalahar, B. K.; Williams, L. D. *Molecular Biology and Evolution* **2009**, *26*, 2415.

65. Sarver, M.; Zirbel, C. L.; Stombaugh, J.; Mokdad, A.; Leontis, N. B. *Journal of mathematical biology* **2008**, *56*, 215.

66. Petrov, A. S.; Bernier, C. R.; Hershkovitz, E.; Xue, Y.; Waterbury, C. C.; Grover, M. A.; C., H. S.; Hud, N. V.; Wartell, R. M.; Williams, L. D. *Nucleic Acids Res.* **2013**, *41*, 7522.

67. Kemena, C.; Notredame, C. *Bioinformatics* **2009**, *25*, 2455.

68. Martins, E. P.; Hansen, T. F. *American Naturalist* **1997**, *149*, 646.

69. Pei, J. *Current Opinion in Structural Biology* **2008**, *18*, 382.

70. Morrison, D. A.; Ellis, J. T. *Molecular Biology and Evolution* **1997**, *14*, 428.

71. Kim, J.; Ma, J. *Bioinformatics* **2014**, *30*, 1010.

72. Chang, J.-M.; Di Tommaso, P.; Notredame, C. *Molecular Biology and Evolution* **2014**, *31*, 1625.

73. Notredame, C. *PLoS Comput Biol* **2007**, *3*, e123.

74. Sahraeian, S. M. E.; Yoon, B.-J. *Nucleic Acids Research* **2011**, *39*, W8.

75. Di Tommaso, P.; Moretti, S.; Xenarios, I.; Orobitg, M.; Montanyola, A.; Chang, J.-M.; Taly, J.-F.; Notredame, C. *Nucleic Acids Research* **2011**, *39*, W13.

76. Havgaard, J. H.; Torarinsson, E.; Gorodkin, J. *PLoS Comput Biol* **2007**, *3*, e193.

77. Gardner, P. P.; Wilm, A.; Washietl, S. *Nucleic Acids Research* **2005**, *33*, 2433.

78. Mallatt, J.; Craig, C. W.; Yoder, M. J. *Mol. Phylogenet. Evol.* **2012**, *64*, 603.

79. Shannon, C. E. *Bell System Technical Journal* **1948**, *27*, 379.

80. Petrov, A. I.; Zirbel, C. L.; Leontis, N. B. *RNA* **2013**, *19*, 1327.

81. ; Schrödinger, LLC. The PyMOL Molecular Graphics System. Version 1.2r3pre ed.

82. Goertzen, L. R.; Cannone, J. J.; Gutell, R. R.; Jansen, R. K. *Mol. Phylogenet. Evol.* **2003**, *29*, 216.

83. Grant, G. R.; Farkas, M. H.; Pizarro, A. D.; Lahens, N. F.; Schug, J.; Brunk, B. P.; Stoeckert, C. J.; Hogenesch, J. B.; Pierce, E. A. *Bioinformatics* **2011**, *27*, 2518.

84. Lindner, R.; Friedel, C. C. *PLoS ONE* **2012**, *7*, e52403.

85. Daugelaite, J.; O' Driscoll, A.; Sleator, R. D. *ISRN Biomathematics* **2013**, *2013*, 14.

86. Kawrykow, A.; Roumanis, G.; Kam, A.; Kwak, D.; Leung, C.; Wu, C.; Zarour, E.; Sarmenta, L.; Blanchette, M.; Waldispühl, J.; Phylo, p. *PLoS ONE* **2012**, *7*, e31362.

87. http://www.arb-silva.de/ 2014.

88. http://www.arb-silva.de/documentation/ In *SILVA Documentation* 2014.

89. Leontis, N. B.; Westhof, E. *Journal of Molecular Biology* **1998**, *283*, 571.

90. Hassouna, N.; Mithot, B.; Bachellerie, J.-P. *Nucleic Acids Research* **1984**, *12*, 3563.

91. Woese, C. R.; Magrum, L. J.; Gupta, R.; Siegel, R. B.; Stahl, D. A.; Kop, J.; Crawford, N.; Brosius, J.; Gutell, R.; Hogan, J. J.; Noller, H. F. *Nucleic Acids Res.* **1980**, *8*, 2275.

92. Noller, H. F.; Kop, J.; Wheaton, V.; Brosius, J.; Gutell, R. R.; Kopylov, A. M.; Dohme, F.; Herr, W.; Stahl, D. A.; Gupta, R.; Waese, C. R. *Nucleic Acids Res.* **1981**, *9*, 6167.

93. Sloof, P.; Van den Burg, J.; Voogd, A.; Benne, R.; Agostinelli, M.; Borst, P.; Gutell, R.; Noller, H. *Nucleic Acids Res* **1985**, *13*, 4171.

94. Woese, C. R.; Gutell, R. R. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, *86*, 3119.

95. Gutell, R. R.; Woese, C. R. *Proceedings of the National Academy of Sciences of the United States of America* **1990**, *87*, 663.

96. Woese, C. R.; Winker, S.; Gutell, R. R. *Proc. Natl. Acad. Sci. U. S. A.* **1990**, *87*, 8467.

97. Gutell, R. R.; Larsen, N.; Woese, C. R. *Microbiol Rev* **1994**, *58*, 10.

98. Ben-Shem, A.; Jenner, L.; Yusupova, G.; Yusupov, M. *Science* **2010**, *330*, 1203.

99. Melnikov, S.; Ben-Shem, A.; Garreau de Loubresse, N.; Jenner, L.; Yusupova, G.; Yusupov, M. *Nature structural & molecular biology* **2012**, *19*, 560.

100. Petrov, A. S.; Bernier, C. R.; Hsiao, C.; Norris, A. M.; Kovacs, N. A.; Waterbury, C. C.; Stepanov, V. G.; Harvey, S. C.; Fox, G. E.; Wartell, R. M.; Hud, N. V.; Williams, L. D. *Proceedings of the National Academy of Sciences* **2014**.

101. Gerbi, S. A. *Biosystems* **1986**, *19*, 247.

102. Klinge, S.; Voigts-Hoffmann, F.; Leibundgut, M.; Arpagaus, S.; Ban, N. *Science* **2011**, *334*, 941.

103. Woese, C. R.; Kandler, O.; Wheelis, M. L. *Proc Natl Acad Sci U S A* **1990**, *87*, 4576.

104. Gribaldo, S.; Brochier-Armanet, C. *Philos Trans R Soc Lond B Biol Sci* **2006**, *361*, 1007.

105. Harish, A.; Tunlid, A.; Kurland, C. G. *Biochimie* **2013**, *95*, 1593.

106. Hahn, J.; Haug, P. *Systematic and Applied Microbiology* **1986**, *7*, 178.

107. Glansdorff, N.; Xu, Y.; Labedan, B. *Biol Direct* **2008**, *3*, 29.

108. Doolittle, W. F. *Sci Am* **2000**, *282*, 90.

109. Fox, G. E. *Cold Spring Harb Perspect Biol* **2010**, *2*, a003483.

110. Agmon, I.; Bashan, A.; Yonath, A. *Isr. J. Ecol. Evol.* **2006**, *52*, 359.

111. Fang, H.; Oates, M. E.; Pethica, R. B.; Greenwood, J. M.; Sardar, A. J.; Rackham, O. J. L.; Donoghue, P. C. J.; Stamatakis, A.; de Lima Morais, D. A.; Gough, J. *Sci. Rep.* **2013**, *3*.

112. Letunic, I.; Bork, P. *Nucleic Acids Research* **2011**, *39*, W475.

113. Hedges, S. B.; Kumar, S. *The timetree of life*; Oxford University Press, 2009.

114. Hedges, S. B.; Dudley, J.; Kumar, S. *Bioinformatics* **2006**, *22*, 2971.

115. Watson, J. D.; Crick, F. H. *Nature* **1953**, *171*, 737.

116. Zuckerkandl, E.; Pauling, L. *J Theor Biol* **1965**, *8*, 357.

117. Woese, C. R. *The genetic code: the molecular basis for genetic expression*; Harper & Row: N.Y., 1967.

118. Crick, F. *J Mol Biol* **1968**, *38*, 367

119. Woese, C. R.; Harper & Row: 1968.

120. Crick, F. *Nature* **1970**, *226*, 561.

121. Miller, S. L.; Orgel, L. E. *The origins of life on the earth*; Prentice-Hall, 1974.

122. Crick, F.; Brenner, S.; Klug, A.; Pieczenik, G. *Orig Life* **1976**, *7*, 389

123. Fox, S. W.; Dose, K. *Molecular evolution and the origin of life*; M. Dekker, 1977.

124. Woese, C. R. *RNA* **2001**, *7*, 1055.

125. Fox, G. E.; Ashinikumar, K. N. In *The Genetic Code and the Origin of Life*; de Pouplana, L. R., Ed.; Kluwer Academic / Plenum Publishers, New York 2004, p 92.

126. Smith, T. F.; Lee, J. C.; Gutell, R. R.; Hartman, H. *Biol Direct.* **2008**, *3*, 16.

127. Caetano-anollés, G.; Wang, M.; Caetano-anollés, D.; Mittenthal, J. E. *Biochem J* **2009**, *417*, 621.

128. Noller, H. F. *Cold Spring Harb Perspect Biol* **2010**, *7*, 7.

129. ; Atkins, J. F., Gesteland, R. F., Cech, T. R., Eds.; Cold Spring Harbor Laboratory Press.

130. Gilbert, W. *Nature* **1986**, *319*, 618.

131. Cech, T. R. *Structure* **1995**, *3*, 969.

132. Bernhardt, H. S.; Tate, W. P. *Biol Direct* **2010**, *5*, 16.

133. Klein, D. J.; Schmeing, T. M.; Moore, P. B.; Steitz, T. A. *EMBO J* **2001**, *20*, 4214.

134. Mears, J. A.; Cannone, J. J.; Stagg, S. M.; Gutell, R. R.; Agrawal, R. K.; Harvey, S. C. *J Mol Biol.* **2002**, *321*, 215.

135. Yonath, A. *Mol Cells.* **2005**, *20*, 1.

136. Agmon, F.; Bashan, A.; Zarivach, R.; Yonath, A. *Biol. Chem.* **2005**, *386*, 833.

137. Hury, J.; Nagaswamy, U.; Larios-Sanz, M.; Fox, G. E. *Orig Life Evol Biosph* **2006**, *36*, 421.

138. Bokov, K.; Steinberg, S. V. *Nature* **2009**, *457*, 977.

139. Leontis, N. B.; Lescoute, A.; Westhof, E. *Curr Opin Struct Biol* **2006**, *16*, 279.

140. Woese, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8742.

141. Fox, G. E.; Tran, Q.; Yonath, A. *Astrobiology* **2012**, *12*, 57.

142. Valle, M.; Zavialov, A.; Sengupta, J.; Rawat, U.; Ehrenberg, M.; Frank, J. *Cell* **2003**, *114*, 123.

143. Lancaster, L.; Lambert, N. J.; Maklan, E. J.; Horan, L. H.; Noller, H. F. *RNA* **2008**, *14*, 1999.

144. Lazcano, A.; Miller, S. L. *The origin and early evolution of life: prebiotic chemistry, the Pre-RNA world, and time*; MIT Press, 1996.

145. Shapiro, R. *Proc Natl Acad Sci U S A* **1999**, *96*, 4396.

146. Orgel, L. E. *Critical Reviews in Biochemistry and Molecular Biology* **2004**, *39*, 99.

147. Costanzo, G.; Pino, S.; Ciciriello, F.; Di Mauro, E. *J Biol Chem* **2009**, *284*, 33206.

148. Johnson, A.; Cleaves, H. J.; Bada, J. L.; Lazcano, A. *Origins of Life and Evolution of Biospheres* **2009**, *39*, 240.

149. Cleaves Ii, H. J. *Journal of Theoretical Biology* **2010**, *263*, 490.

150. Benner, S. A.; Kim, H. J.; Carrigan, M. A. *Acc Chem Res* **2012**.

151. Holm, N. G. *Geobiology* **2012**.

152. Hud, N. V.; Lynn, D. G. *Curr Opin Chem Biol* **2004**, *8*, 627.

153. Engelhart, A. E.; Cafferty, B. J.; Okafor, C. D.; Chen, M. C.; Williams, L. D.; Lynn, D. G.; Hud, N. V. *ChemBioChem* **2012**, *13*, 1121.

154. Hsiao, C.; Chou, I. C.; Okafor, C. D.; Bowman, J. C.; O'Neill, E. B.; Athavale, S. S.; Petrov, A. S.; Hud, N. V.; Wartell, R. M.; Harvey, S. C.; Williams, L. D. *Nat Chem* **2013**, *5*, 525.

155. Kavran, J. M.; Steitz, T. A. *Journal of Molecular Biology* **2007**, *371*, 1047.

156. Dinman, J. D. *International journal of biomedical science : IJBS* **2005**, *1*, 2.

157. Bock, L. V.; Blau, C.; Schroder, G. F.; Davydov, II; Fischer, N.; Stark, H.; Rodnina, M. V.; Vaiana, A. C.; Grubmuller, H. *Nature structural & molecular biology* **2013**, *20*, 1390.

158. Athavale, S. S.; Gossett, J. J.; Hsiao, C.; Bowman, J. C.; O'Neill, E.; Hershkovitz, E.; Preeprem, T.; Hud, N. V.; Wartell, R. M.; Harvey, S. C.; Williams, L. D. *RNA* **2012**, *18*, 752.

159. Wolf, Y. I.; Aravind, L.; Grishin, N. V.; Koonin, E. V. *Genome Res* **1999**, *9*, 689.

160. Cavalier-Smith, T. *Biol Direct.* **2006**, *1*, 19.

161. Krupkin, M.; Matzov, D.; Tang, H.; Metz, M.; Kalaora, R.; Belousoff, M. J.; Zimmerman, E.; Bashan, A.; Yonath, A. *Philos Trans R Soc Lond B Biol Sci* **2011**, *366*, 2972.

162. Belousoff, M. J.; Davidovich, C.; Zimmerman, E.; Caspi, Y.; Wekselman, I.; Rozenszajn, L.; Shapira, T.; Sade-Falk, O.; Taha, L.; Bashan, A.; Weiss, M. S.; Yonath, A. *Biochem Soc Trans* **2010**, *38*, 422.